Hochschule Hannover

Fakultät III – Medien, Information und Design

Abteilung Information und Kommunikation

Automatische Klassifizierung medizinischer Literatur durch Analyse verfügbarer Notationen

Bachelorarbeit

im Studiengang Informationsmanagement

vorgelegt von

Andreas Lüschow

Erstgutachter: Prof. Dr. Christian Wartena

Zweitgutachterin: Dr. Ina Blümel

Hannover, den 10.12.2016

Abstract

In den letzten Jahren ist, nicht zuletzt aufgrund der schnellen und einfachen Verfügbarkeit von Daten und Informationen, ein Anstieg an veröffentlichter Literatur zu beobachten. Bibliotheken stehen vor der Herausforderung, diese Ressourcen zu erschließen und damit verfügbar zu machen. Ein Teilaspekt ist hierbei die Klassifizierung. Die Arbeit untersucht Voraussetzungen und Möglichkeiten der automatischen Klassifizierung am Beispiel medizinischer Literatur. Der erste, theoretische Teil beinhaltet die Beschreibung der Grundlagen der Inhaltserschließung, des Data Mining und der automatischen Klassifizierung sowie eine umfassende Übersicht über den aktuellen Forschungsstand in diesem Bereich. Im zweiten Teil wird die Auswahl, Aufbereitung und Analyse eines aus Katalogdatensätzen der Bibliothek der Medizinischen Hochschule Hannover bestehenden Datenbestandes erläutert. Die Anwendung von Verfahren des maschinellen Lernens zur Klassifizierung bibliographischer Datensätze wird am Beispiel des Algorithmus *k-nearest-neighbours* verdeutlicht. Hierbei lässt sich eine korrekte Klassifizierung von rund 58 % der Dokumente erreichen. Abschließend werden Optimierungsansätze (z.B. semi-automatische Verfahren) und Herausforderungen automatischer Klassifizierungsverfahren (z.B. uneinheitlich erschlossene Datensätze oder ungleiche Verteilung der Klassen einer Systematik in den Dokumenten) aufgezeigt.

Inhaltsverzeichnis

Teil 2 - Untersuchungen

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

Um Dokumente auffindbar (und somit nutzbar) zu machen, müssen diese durch Metadaten erschlossen werden. Das Ergebnis dieser Erschließung findet sich dann z.B. als Katalogdatensatz im Online-Katalog (OPAC) einer wissenschaftlichen Bibliothek wieder. Hierbei werden zwei Teilbereiche unterschieden: *Formalerfassung* und *Inhaltserschließung*. Zur formalen Erfassung gehören u.a. Angaben über Titel, Autor(en) und ISBN einer Publikation sowie der Verlagsname und das Erscheinungsjahr. Zu den inhaltlichen Erschließungselementen zählen u.a. Klassifikationen; im Laufe der Zeit haben sich an unterschiedlichen Orten und für unterschiedliche Zwecke verschiedene Klassifikationen herausgebildet.

Während meines ersten Praktikums, das in der Bibliothek der Medizinischen Hochschule Hannover (MHH) stattfand, bin ich durch eine Projektarbeit mit dem Themenbereich Systematiken/Klassifikationen in Berührung gekommen. Das Ziel war damals, eine OPAC-Suchmöglichkeit anhand der Klassen der Klassifikation der *National Library of Medicine* (NLM) zu ermöglichen. Diese wird in der MHH für die Aufstellung der Freihandbestände verwendet. Mithilfe dieses Sucheinstiegs sollen die Nutzer einen Überblick über alle vorhandenen Ressourcen aus dem für sie relevanten Klassifikationszweig erhalten können.

Problematisch war damals, dass nicht alle Datensätze mit der NLM-Klassifikation erschlossen sind. Insbesondere E-Books, deren Daten automatisch z.B. von den Verlagen eingespielt werden, verfügen oft nicht über die benötigten Metadaten. Allerdings sind häufig bereits Metadaten anderer Klassifikationen vorhanden. Da die Menge der lizenzierten E-Books bei Weitem die Möglichkeiten der manuellen Sacherschließung übersteigt, werden diese Datensätze nicht über die Verlagsdaten hinaus erschlossen. Dementsprechend sind sie auch über die in der Projektarbeit entwickelten Suche anhand der Systematik nicht auffindbar.

Die Bachelorarbeit soll auf diese Projektarbeit aufbauend nun untersuchen, inwieweit sich die in den Datensätzen des OPAC der Bibliothek der MHH vorhanden Klassifikationen aufeinander abbilden lassen, um zukünftig eventuell durch (semi-)automatische Verfahren anhand der vorliegenden Klassifikationen eines Datensatzes die passende NLM-Klassifikation bestimmen zu können.

Fragestellung und Zielsetzung

Die Auswahl des Themas der Bachelorarbeit basiert auf der Vermutung, dass sich die Klassen der einzelnen Klassifikationen befriedigend einander zuordnen lassen. Hierdurch sollen

die Metadaten der bislang nicht oder nur unvollständig erschlossenen Datensätze angereichert werden können. Die Bachelorarbeit soll nun eine Antwort auf die Frage finden, ob und in welchem Maße die NLM-Klassifikation durch maschinelle Auswertung vorhandener anderer Klassifikationen ermittelt werden kann. Als Ergebnis dieser Untersuchung sollen Möglichkeiten und Probleme formuliert werden, die zukünftig bei der Konzeptionierung automatischer Sacherschließungsverfahren Berücksichtigung finden können.

Diese Ergebnisse können dazu beitragen, die Chancen und Grenzen zukünftiger automatischer Sacherschließungsvorhaben zu bewerten. Dies bleibt allerdings aufgrund des Umfangs der Arbeit und der Beschränkung auf nur eine Klassifikation auf den Bereich der medizinischen Bibliotheken beschränkt. Des Weiteren ermöglichen die durchgeführten Analysen der NLM-Klassifikation den Mitarbeitern medizinischer Bibliotheken (und hier insbesondere der MHH), ihre bisher angewandten Sacherschließungsmethoden zu hinterfragen. Eine Klassifizierung unerschlossener Datensätze durch automatisch ermittelte oder dem Katalogisierer zumindest vorgeschlagene bzw. empfohlene NLM-Klassen könnte die Arbeit der Sacherschließung effizienter und zeitsparender gestalten.

Gliederung

Die Arbeit gliedert sich in zwei Teile: Im ersten Teil werden die theoretischen Grundlagen für die im zweiten Teil durchgeführten Untersuchungen erläutert. Zunächst werden die relevanten Aspekte der Themen *Inhaltserschließung* und *Maschinelles Lernen* erklärt. Im Anschluss daran findet sich ein ausführlicher Abschnitt über das diese beiden Bereiche beinhaltende Forschungsfeld der *automatischen Klassifizierung*. Danach wird der Forschungsstand auf diesem Feld beschrieben. Der erste Teil schließt mit der Vorstellung derjenigen bibliothekarischen Systematiken ab, die im zweiten Teil untersucht werden.

Der zweite Teil der Arbeit erläutert den Aufbau und die Ergebnisse der im Rahmen dieser Arbeit durchgeführten Untersuchungen. Zunächst wird der den Untersuchungen zugrundeliegende Datenbestand beschrieben. Die Aufbereitung und Analyse der Daten folgt in einem weiteren Schritt. Anschließend wird unter Verwendung der im ersten Teil der Arbeit vorgestellten Methoden der Datenbestand untersucht. Zuletzt werden die hierbei ermittelten Ergebnisse zusammengefasst und ihre Relevanz für die Fragestellung dieser Arbeit diskutiert. Ein Ausblick auf weitere nötige Forschungen beendet die Arbeit.

TEIL 1 - GRUNDLAGEN

Im ersten Teil dieser Arbeit werden die im weiteren Verlauf genutzten wesentlichen Begriffe und Themen definiert, in ihren jeweiligen (thematischen) Zusammenhang eingeordnet und zueinander abgegrenzt. Hierdurch wird nicht nur ein einheitliches Verständnis der später beschriebenen Sachverhalte erreicht, sondern gleichzeitig ein Überblick über die in der Literatur vorrangig verwendeten Bezeichnungen gegeben. Da für die Erstellung dieser Arbeit zu etwa gleichen Teilen deutsch- und englischsprachige Literatur verwendet wurde, dient die folgende Begriffsklärung auch der Vereinheitlichung der in diesen beiden Sprachen oftmals unterschiedlichen Bezeichnungen.

1 Inhaltliche Erschließung

Indexierung bezeichnet die „Repräsentation von Dokumenteninhalten über Metainformationen“[1] mit dem „Ziel der Inhaltserschließung und der gezielten Wiederauffindung“[2]. Klassischerweise werden folgende Arbeitsschritte zum Indexieren gezählt:

a) Die Inhaltsanalyse, bei welcher der Inhalt des Dokuments begrifflich zu erfassen versucht wird.
b) Die anschließende Inhaltsrepräsentation durch die Elemente einer Indexierungssprache. Im Zusammenhang mit Klassifikationssystemen werden hierunter die Notationen verstanden.[3]

Indexierung und inhaltliche Erschließung sind schwer voneinander abzugrenzen, sodass in der vorliegenden Arbeit beide Begriffe quasi-synonym verstanden werden.

Grundlage der inhaltlichen Erschließung sind *Dokumente*. Die Fachliteratur unterscheidet zwar zwischen *Dokument, dokumentarischer Bezugseinheit* und *Dokumentationseinheit*[4], eine solche – in anderen Bereichen sicherlich notwendige – Differenzierung ist für die hier durchgeführten Untersuchungen jedoch nicht zielführend. Nach ISO 5963:1 versteht man unter einem Dokument einen Informationsträger, der einer inhaltlichen Erschließung zugänglich ist.[5] Somit zählen auch (Meta-)Datensätze oder (HTML-)Webseiten als Dokument. Diese allgemeine Definition wird in dieser Arbeit verwendet.

[1] Nohr (2005), S. 21
[2] Ebd., S. 24
[3] Vgl. ebd., S. 24f.
[4] Vgl. u.a. Bertram (2005), S. 21
[5] Vgl. ebd.

Ziel inhaltlicher Erschließung ist u.a. ein verbessertes *Information Retrieval*, dessen Gegenstand „die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen“[6] ist. Es ist als ein technisch gestützter Prozess zu verstehen, der den Wissenstransfer vom Wissensproduzenten zum Informations-Nachfragenden zum Ziel hat.[7] Klassische Anwendungen des Information Retrieval sind z.B. Bibliothekskataloge.

Die inhaltliche Erschließung produziert *Metadaten*, also „Daten über Daten“. Diese dienen der Beschreibung von Dokumenten und schaffen einen Mehrwert, indem sie nicht nur Informationen aus den Dokumenten selbst beinhalten (wie z.B. den Titel, Autoren und das Veröffentlichungsjahr), sondern auch Metainformationen der Dokumente berücksichtigen. Hierzu zählen z.B. Einordnungen in Klassifikationssysteme und Inhaltsbeschreibungen durch Schlagwörter.

Eine *Klassifikation* ist „im Kontext der Information und Dokumentation (…) eine künstlichsprachige Dokumentationssprache zur inhaltlichen Groberschließung. Sie ist ein Begriffssystem, das zur Ordnung von Gegenständen oder Wissen über Gegenstände eingesetzt wird und auf dem Prinzip der Klassenbildung beruht.“[8] Einer Klasse innerhalb einer Klassifikation werden Elemente zugeordnet, die mindestens in einem Merkmal übereinstimmen und somit in einem verwandtschaftlichen Verhältnis stehen. Weiterhin kann es Oberklassen und Unterklassen geben, die zueinander in einem hierarchischen Verhältnis stehen. Im bibliothekarischen Bereich wird der Begriff *Systematik* synonym zum Begriff Klassifikation verwendet.

Der Vorgang der Zuordnung eines Elementes zu einer Klasse wird als *Klassieren* bezeichnet. Unter *Klassifizieren* versteht man hingegen den Prozess der Klassenbildung, bei dem die eine Klasse definierenden Merkmale festgelegt werden.[9] In der englischsprachigen Literatur finden sich die Begriffe *to class* bzw. *to classify*. Beide Begriffe werden in der Literatur oftmals synonym verwendet. Sebastiani spricht von „*Text categorization* (TC – a.k.a. *text classification*, or *topic spotting*)“ als „activity of labeling natural language texts with thematic categories from a predefined set.“[10] Im Folgenden wird also unter *Klassifizieren* der Prozess der Zuordnung eines Dokumentes zu einer existierenden Klasse verstanden. Die nur

[6] Nohr (2005), S. 19
[7] Vgl. Knorz (1995), zit. nach Nohr (2005), S. 19
[8] Bertram (2005), S. 150
[9] Vgl. ebd.
[10] Sebastiani (2002), S. 1

am Rande erwähnte Erstellung von Klassen wird im Zusammenhang mit dem Begriff des *Clustering* beschrieben.

Der Begriff *automatische Inhaltserschließung* beinhaltet Verfahren, die durch maschinelle (d.h. in der Regel durch Computer durchgeführte) Automatismen inhaltsanalytische Ergebnisse produzieren. Konkret werden das *automatische Abstracting* und das *automatische Indexieren* hierunter zusammengefasst. Im Gegensatz zur intellektuell durchgeführten Inhaltserschließung werden bei der automatischen Inhalterschließung jedoch nicht zwingend für den Nutzer sichtbare Metadaten produziert.[11] So können z.B. auch für die Datenverarbeitung im Zusammenhang mit dem maschinellen Lernen benötigte Zwischenstufen berechnet werden.

Als Teil der automatischen Inhaltserschließung umfasst das *automatische Indexieren* alle Verfahren, die auf automatische Weise Dokumente analysieren und daraufhin bestimmte Terme aus diesen Dokumenten extrahieren oder sie bereits vorhandenen kontrollierten Indexierungssprachen wie Thesauri oder Klassifikationen zuweisen.[12] In der vorliegenden Arbeit wird lediglich auf einen Teilbereich des automatischen Indexierens – das *automatische Klassifizieren* – eingegangen.

Nohr spricht von *automatischer Klassifikation*, wenn eine Klassifikation selbst das Ergebnis eines automatischen Verfahrens darstellt.[13] Dies wird in aller Regel durch *Clustering* erreicht. In englischsprachiger Literatur werden „automatic classification" und „clustering" oft als identisch betrachtet.[14] Die Zuordnung von Dokumenten zu bereits bestehenden Klassifikationen entspricht dem *automatischen Klassifizieren*. Da das Clustering in der vorliegenden Arbeit nur am Rande thematisiert wird, wird auch nicht näher auf die Verfahren der automatischen Klassifikation eingegangen.

2 Data Mining und maschinelles Lernen

2.1 Definitionen

Eine umfassende und zufriedenstellende Definition von *Data Mining* ist schwierig. „It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term *data mining* does not really present all the major components in the

[11] Vgl. Bertram (2005), S. 97f.
[12] Vgl. Nohr (2005), S. 27
[13] Vgl. ebd., S. 37, Fußnote
[14] Vgl. z.B. Han u.a. (2012), S. 445

picture.“[15] Eine kurze, prägnante Definition findet sich bei Cleve u.a.: „Data Mining (Datenschürfen) ist die Extraktion von Wissen aus Daten“[16], mit der Einschränkung, dass dieses Wissen vorher unbekannt war und nützlich ist. Außerdem sollte der Prozess der Wissensgewinnung weitgehend automatisch ablaufen. Han betont die Prozesshaftigkeit des Data Mining und dass es sich um große Datenmengen („*large* amount of data“) handelt.[17]

Man unterscheidet zwischen *unstrukturierten*, *semi-strukturierten* und *strukturierten Daten*. Unstrukturierte Daten finden sich z.B. in Form von Texten und Bildern, semi-strukturierte Daten in Form von Webseiten und strukturierte Daten in Datenbanken oder Datenformaten mit einer festgelegten Struktur. Anhand dieser drei Begriffe lässt sich Data Mining zu den angrenzenden Disziplinen *Text Mining* und *Web Mining* abgrenzen: Während bei ersterem vorwiegend unstrukturierte Daten betrachtet werden, sind die Daten im Web Mining in aller Regel durch Verwendung von Auszeichnungssprachen wie HTML semi-strukturiert.[18]

Data Mining benutzt Methoden aus unterschiedlichen Wissenschaftsbereichen. Für die Untersuchungen im zweiten Teil dieser Arbeit sind vor allem die *Statistik* und das *maschinelle Lernen* von Bedeutung. Oftmals lassen sich schon mithilfe statistischer Methoden ausreichend aussagekräftige Ergebnisse erzielen, sodass nicht immer die Anwendung von Data-Mining-Methoden vonnöten ist.[19]

Die auch für einen Computer einfachste Strategie des Lernens ist das Auswendiglernen. Hierbei wird das in den Eingabedaten enthaltene Wissen abgespeichert, sodass es bei Bedarf wieder abgerufen werden kann. Im Kontext der Künstlichen Intelligenz versteht man unter *maschinellem Lernen* jedoch das Erkennen von Mustern und Zusammenhängen in den Eingabedaten. Diese Zusammenhänge können dann auf weitere Daten angewendet werden, um bislang unentdecktes Wissen in diesen zu erkennen.[20] (Maschinelles) Lernen lässt sich wie folgt definieren: „Things learn when they change their behavior in a way that makes them perform better in the future.“[21] Hier wird die Bedeutung der stetigen Leistungsverbesserung hervorgehoben.

In der Literatur findet sich ebenfalls der Begriff *Knowledge Discovery from Data (KDD)*. KDD bezieht sich vor allem auf den gesamten Prozess der Wissensgewinnung (inklusive

[15] Ebd., S. 5
[16] Cleve u.a. (2016), S. 38
[17] Vgl. Han u.a. (2012), S. 8
[18] Vgl. Cleve u.a. (2016), S. 37-39
[19] Vgl. ebd., S. 14
[20] Vgl. ebd.
[21] Witten u.a. (2011), S. 7

notwendiger Schritte wie z.B. der Datenbereinigung), während Data Mining im engeren Sinne nur die eigentliche Analyse der Daten bezeichnet. Da KDD in der Literatur in der Regel synonym zu Data Mining gebraucht wird, wird in der vorliegenden Arbeit auf diese Unterscheidung ebenfalls verzichtet.

2.2 Ausgangsdaten

Eine Datenmenge (z.B. die mithilfe einer Abfrage entstandene Untermenge einer bibliographischen Datenbank), die mit Data-Mining-Verfahren untersucht werden soll, besteht aus mehreren Datensätzen. Jeder dieser Datensätze kann als ein individuelles Objekt betrachtet werden. Dieses Objekt wird wiederum durch mehrere Attribute charakterisiert. Attribute geben also die Eigenschaften eines Objektes wieder. In einer bibliographischen Datenbank repräsentieren die einzelnen Datensätze z.B. Bücher aus dem Bestand einer Bibliothek. Für jeden dieser Datensätze sind wiederum mehrere Attribute, wie z.B. Titel, Autor oder Erscheinungsjahr angegeben. Alle Datensätze besitzen dieselben Attribute, nur deren jeweilige Ausprägung in einem konkreten Wert kann unterschiedlich sein. Wenn ein Datensatz für ein bestimmtes Attribut keinen Wert aufweist, bleibt das entsprechende Datenfeld leer.

Je nach Struktur der Datenmenge können die Datensätze auch ein oder mehrere Attribute besitzen, die den Datensatz einer Klasse oder einem Konzept zuordnen. In einer bibliographischen Datenbank existieren in der Regel für die einzelnen Werke Zuordnungen zu Klassifikationen. Somit lassen sich z.B. alle Veröffentlichungen zusammenfassen, die sich mit dem Thema „Anatomie“ beschäftigen, indem innerhalb des Datenbestandes nur diejenigen Datensätze ausgewählt werden, welche als Attributwert an der entsprechenden Stelle einen bestimmten Wert aufweisen.[22]

Um ein erstes Verständnis für die Zusammensetzung und Struktur der zu den einzelnen in der Datenmenge vorhandenen Klassen gehörenden Datensätze zu bekommen, werden die Klassen beschrieben. Diese Beschreibung kann auf zwei Wegen geschehen[23]:

a) Durch *data characterization*: In diesem Fall werden die wesentlichen Eigenschaften der in diese Klasse gehörenden Datensätze zusammengefasst. Hierfür werden oft statistische Verfahren und Visualisierungen verwendet. So lässt sich z.B. mittels eines Säulendiagramms leicht die Verteilung von Büchern auf die einzelnen Klassen einer Klassifikation darstellen.

[22] Die in diesem Abschnitt beschriebenen Aspekte finden sich in jeder Einführung zum Thema Data Mining. Vgl. z.B. Han u.a. (2012), Cleve u.a. (2016) oder Witten u.a. (2011).
[23] Vgl. Han u.a. (2012), S. 15f.

b) Durch *data discrimination*: Hierbei werden die Eigenschaften einer Klasse mit denen einer oder mehrerer anderer Klassen verglichen. Die Darstellung der Ergebnisse geschieht auf dieselbe Weise wie bei der data characterization. Besonderes Augenmerk sollte auf Regeln und Zusammenhänge gelegt werden, die beim Unterscheiden der untersuchten Klassen helfen. Ein Resultat der data discrimination könnte z.B. sein, dass für zwei ausgewählte Klassen einer bibliothekarischen Klassifikation für jedes Jahr seit 1975 die Anzahl der in diesen beiden Klassen neu hinzugefügten Datensätze dargestellt wird. So könnte festgestellt werden, ob sich die Verteilung der Werke auf die beiden Klassen mit der Zeit geändert hat, woraus ein verändertes Katalogisierungsverhalten abgeleitet werden könnte.

2.3 Unüberwachtes und überwachtes Lernen

Aus vorgegebenen Daten extrahierte Muster können mit zwei Zielsetzungen gewonnen werden: entweder rein zur Beschreibung der Daten oder zur Vorhersage bei Anwendung der Muster auf neue Daten.[24] Letzteres beinhaltet ersteres, da jedes Entdecken von Mustern erst einmal mit der beschreibenden Analyse der Daten beginnen muss. Neben diesen beiden Motivationen, Data-Mining-Verfahren anzuwenden, lassen sich bei der Durchführung derselben zwei Ansätze unterscheiden: *unüberwachtes Lernen* und *überwachtes Lernen*.[25]

2.3.1 Unüberwachtes Lernen

Hierbei sind die Ergebnisse der Analyse vollständig unbekannt. Aus den Eingabedaten werden Muster und Zusammenhänge ermittelt, um die vorhandenen Datensätze in Gruppen von ähnlichen Datensätzen zu organisieren. Dies geschieht, ohne dass in den Daten schon Angaben über eine korrekte Zuordnung vorhanden sind. Ein Beispiel für den Ansatz des unüberwachten Lernens ist das *Clustering*. Durch Clustering kann man bisher nicht bekannte Gruppen von Datensätzen aufspüren; also Datensätze, die sich in bestimmten Eigenschaften ähnlich sind. Es findet bei Internetdiensten wie z.B. Suchmaschinen oder Bilderkennungssoftware Verwendung. Mithilfe von Clustering lassen sich z.B. Dokumente zu einem gemeinsamen Thema ermitteln, ohne dass diese einzeln intellektuell erschlossen werden müssen.[26] Da die Eingabedaten nicht mit einer vorgegebenen Klasse erschlossen sind, lässt sich die

[24] Vgl. ebd.
[25] Vgl. für die folgenden Ausführungen z.B. Cleve u.a. (2016), S. 55
[26] Vgl. Han u.a. (2012), S. 444

Güte der maschinellen Zuordnung allerdings auch nicht messen. Im zweiten Teil dieser Arbeit wird nicht mit Clustering-Methoden gearbeitet, sodass sich die Ausführungen an dieser Stelle auf diesen groben Überblick beschränken.

2.3.2 Überwachtes Lernen

Wenn in den zu untersuchenden Datensätzen Attribute gegeben sind, welche die Zuordnung zu einer bestimmten Klasse vorgeben, spricht man vom überwachten Lernen. Hierbei sind die Dokumente also schon manuell erschlossen und kategorisiert. In einem ersten Schritt erstellt dann ein Lernprozess automatisch einen sogenannten *Klassifikator* für jede vorhandene Klasse. Dies geschieht durch Analyse der Eigenschaften der dieser Klasse zugeordneten Dokumente. Dadurch werden Merkmale ermittelt, welche ein neu zu klassifizierendes Dokument aufweisen sollte, um der betreffenden Klasse zugeordnet zu werden. Da die bereits klassifizierten Dokumente (neben dem den Klassifikator erstellenden Lernprozess) die wichtigste Komponente dieses Verfahrens darstellen, kommt ihrer Aufbereitung und Analyse tragende Bedeutung zu.[27] Durch die in den Daten bereits vorgegebene korrekte Zuordnung lässt sich beim überwachten Lernen (im Unterschied zu den Verfahren des unüberwachten Lernens) die Qualität des verwendeten Data-Mining-Verfahrens messen.

2.4 Phasen des Data Mining

Es werden fünf aufeinander aufbauende Phasen im Data-Mining-Prozess unterschieden.[28] In jeder Phase werden für den weiteren Verlauf wichtige Entscheidungen getroffen, sodass eine gründliche und durchdachte Bearbeitung der einzelnen Schritte wesentlich zum Erfolg des Prozesses beiträgt. Die fünf Phasen sind:

2.4.1 Selektion

In dieser Phase findet die Auswahl der für die Analyse geeigneten und benötigten Datenmengen statt. Wichtige Voraussetzung ist, im Vorhinein die Verfügbarkeit der betreffenden Daten zu klären. Gerade Aufgaben, die sich nicht mit Daten aus frei zugänglichen Quellen bearbeiten lassen, erfordern oftmals intensive Recherchen oder z.B. einen Zugang zu firmeneigenen Datenbeständen. Daher ist eine genaue Fragestellung und Aufgabenbeschreibung notwendig, um Klarheit über den Umfang und die Struktur der benötigten Daten zu bekommen. Bereits in dieser ersten Phase werden unter Umständen für den Erfolg der Untersuchungen wichtige Entscheidungen getroffen, da die nachträgliche Beschaffung von Daten und Informationen zeit- und kostenintensiv sein kann. Fällt z.B. während einer späteren

[27] Vgl. Oberhauser (2005), S. 22
[28] Vgl. Fayyad u.a. (1996), dargestellt nach Cleve u.a. (2016), S. 5f. und S. 10-12

Phase auf, dass weitere Attribute der ausgewählten Datensätze benötigt werden, müssen diese nicht nur beschafft, sondern auch in den womöglich schon bereinigten und vorverarbeiteten Datenbestand integriert werden. Bei umfangreichen Datenbeständen ist z.B. darüber hinaus aus Kapazitätsgründen oftmals zu klären, ob eventuell nur eine Teilmenge für die weitere Verarbeitung betrachtet werden soll. Ist dies der Fall, muss bestimmt werden, anhand welcher Kriterien diese Teilmenge möglichst repräsentativ für den gesamten Bestand ausgewählt werden kann.

Durch Im- und Export der für notwendig und geeignet erachteten Daten (aus möglicherweise unterschiedlichsten Quellen) schafft man nun eine einheitliche Datengrundlage. Hierzu werden die Daten in der Regel in eine Datenbank oder in eine Datentabelle übertragen.

2.4.2 Datenvorverarbeitung

Im Anschluss an die Datenauswahl und den Import in den eigenen Datenbestand müssen die Daten bereinigt werden. Hierfür wird zunächst die Qualität des Datenbestandes untersucht und mit den eigenen Anforderungen verglichen. Da sich in bis zu 5 % der Felder eines Datenbestandes unvollständige, fehlerhafte oder sich widersprechende Einträge finden[29], trägt die Datenvorverarbeitung durch Minimalisierung dieser Fehlerquote wesentlich zur Qualität der Untersuchungsergebnisse bei. Weiterhin können Datenfelder mit unpassenden Werten bzw. Datentypen belegt sein oder die Datensätze können beim Import beschädigt worden sein. Häufig wird erst durch die Aufbereitung der Daten eine tiefergehende Analyse ermöglicht, da viele der in den späteren Phasen verwendeten Methoden auf konsistente und saubere Daten angewiesen sind. Außerdem werden die einzelnen hier vorgestellten Phasen nicht zwangsläufig von ein und derselben Person bearbeitet, sodass der spätere Anwender der Analysewerkzeuge auf die Qualität der Daten vertrauen können muss. Insbesondere bei der Verwendung automatisierter Datenanalysen ist eine erhöhte Anfälligkeit gegenüber Datenmängeln gegeben. Während der Datenvorverarbeitung findet also auch immer schon eine erste Analyse statt, wodurch dem Datenbestand innewohnende Mängel und Besonderheiten bereits früh erkannt werden können. Dies trägt zur möglichst optimalen Anpassung der späteren Lernverfahren an den Datenbestand bei.

2.4.3 Transformation

Je nach verwendetem Data-Mining-System und dessen Anforderungen an die Eingangsdaten können unterschiedliche Datenformate benötigt werden. Nach Bereinigung der Datenmenge

[29] Vgl. Cleve u.a. (2016), S. 10

muss diese nun eventuell in ein passendes Format transformiert werden. Dabei können z.B. die Umcodierung von Attributen, deren Normalisierung oder auch die Festlegung von Trennzeichen zwischen den einzelnen Datensätzen nötig sein. Um die folgende Analyse und Auswertung zu erleichtern, ist häufig auch die Übersetzung von in textlicher Form vorhandenen Daten in Codes nötig. Hierbei hängt es mitunter von den verwendeten Lernalgorithmen ab, welche Umformungen sinnvoll sind.

2.4.4 Data Mining

In dieser Phase findet das eigentliche Data Mining statt. Die bisherigen drei Phasen sind die wichtige Vorbereitung für die nun durchgeführten Analysen. Die grundlegende Data-Mining-Aufgabe wird bestimmt (z.B. Clustering oder Klassifizierung) und daran anschließend werden geeignete Verfahren ausgewählt. Parameter und Konfigurationen der Verfahren müssen eingestellt werden, damit die Analyse den eigenen Vorstellungen und Anforderungen entsprechend verläuft. Schließlich erstellt das gewählte Verfahren ein Modell (z.B. einen Entscheidungsbaum) für die Beschreibung der Daten.

2.4.5 Interpretation und Evaluation

Die erzielten Resultate müssen abschließend interpretiert werden. Leitende Fragen können z.B. sein: Sind die Ergebnisse neu oder bereits bekannt? Lassen sie sich auf andere Daten übertragen? Inwiefern spiegeln sie Regelmäßigkeiten wider? Wie lassen sich diese Regelmäßigkeiten erklären?

Weiterhin sollen die Muster den Kriterien der *Gültigkeit*, *Neuartigkeit*, *Nützlichkeit* und *Verständlichkeit* genügen. Dann ist sichergestellt, dass es sich um neues, verwendbares Wissen handelt. Viele im Rahmen der Analysen gefundene Muster entsprechen diesen Kriterien nicht, da sie trivial, bekannt oder für die Fragestellung unbedeutend sind. Die *Gültigkeit* gibt objektiv wieder, mit welcher Sicherheit sich das gefundene Modell auf neue Daten übertragen lässt. *Neuartigkeit* gibt Aufschluss darüber, ob das gefundene Modell den bisherigen Wissensstand erweitert oder korrigiert. Anhand der *Nützlichkeit* wird beurteilt, inwiefern das Modell praktische Relevanz hat. Nicht zu vernachlässigen ist die *Verständlichkeit*, welche die Nachvollziehbarkeit des gefundenen Modells für den Menschen misst. Gerade bei Anwendungsgebieten, die klassischerweise intellektuell bearbeitet werden – wie z.B. der Klassifizierung von Bibliotheksbeständen –, ist es für die Akzeptanz von automatisierten Verfahren wichtig, dass nicht nur deren Ergebnisse, sondern auch die Ermittlung derselben für den Menschen nachvollziehbar dargestellt werden können.

Stellt sich in dieser letzten Phase heraus, dass in den vorhergehenden Phasen getroffene Entscheidungen nicht zu den gewünschten Ergebnissen geführt haben, findet eine erneute Datenbearbeitung in der entsprechenden Phase statt, sodass die Analysen mit neuen Voraussetzungen ein weiteres Mal durchlaufen werden können. Dies kann von wenig umfangreichen Eingriffen wie dem Ignorieren bestimmter Attribute bis hin zur vollständigen Neukonzeptionierung der Vorgehensweise reichen.

Neben dem hier ausführlich vorgestellten Modell gibt es eine Reihe weiterer Modelle, die den Prozess des Data Mining in seinem Ablauf beschreiben und strukturieren. Diese spiegeln oft die spezifischen Anforderungen einer bestimmten Branche oder einer bestimmten Herangehensweise wider. Beispiele sind das CRISP-DM-Modell (Cross Industry Standard Process for Data Mining) oder das SEMMA-Modell (Sample, Explore, Modify, Model and Assess).[30] Bei Han u.a. findet sich ein 7-Phasen-Modell.[31] Auf diese weiteren Modelle soll hier nicht näher eingegangen werden, da sie im Kern alle sehr ähnlich sind. Die Untersuchungen im zweiten Teil der vorliegenden Arbeit orientieren sich weitestgehend am oben beschriebenen Modell nach Fayyad u.a.

3 Automatisches Klassifizieren

Anknüpfend an die oben ausgeführten Begriffsdefinitionen und Erläuterungen wird im folgenden Abschnitt die Methode des automatischen Klassifizierens beschrieben, welche die Grundlage für die im zweiten Teil dieser Arbeit durchgeführten Untersuchungen bildet.

3.1 Begriff

Wie in Abschnitt 1 dargestellt, handelt es sich beim automatischen Klassifizieren um einen Teilbereich der automatischen Indexierung, welche wiederum einen Teilbereich der automatischen Inhaltserschließung darstellt. Die DIN 31623-1 unterscheidet drei Indexierungs- bzw. – im Zusammenhang dieser Arbeit – Klassifizierungsmethoden[32]:

a) Bei der *intellektuellen Klassifizierung* werden die Notationen einer Systematik durch einen Menschen vergeben. Es folgt zunächst eine Inhaltsanalyse des Dokumentes, die ein menschliches Verständnis des Bedeutungsgehalts zum Ziel hat.

[30] Für kurze Beschreibungen dieser beiden alternativen Modelle siehe ebd., S. 6-9
[31] Vgl. Han u.a. (2012), S. 6-8
[32] Vgl. Nohr (2005), S. 27f.

b) Von *computerunterstützter Klassifizierung* spricht man, wenn einem menschlichen Katalogisierer mit Hilfe eines Computers passende Klassen bzw. Notationen – eventuell in einer nach Relevanz berechneten Reihenfolge – vorgeschlagen werden. Dies wird auch als *semi-automatische Klassifizierung* bezeichnet[33], da sich der Entscheidungsprozess des Katalogisierers auf diese maschinell ermittelten Klassen stützt.
c) Die *automatische Klassifizierung* vergibt Notationen, ohne dass eine Kontrolle oder Bestätigung durch einen Menschen stattfindet. In diesem Fall spricht man auch von „harter" Klassifizierung, weil für jede Kombination aus Dokument und Klasse entschieden werden muss, ob das Dokument der Klasse zugehört oder nicht. Dies ist ein wesentlicher Grund, weshalb rein automatische Systeme in der Praxis selten angewendet werden.

Beim Einsatz von Datenverarbeitungsanlagen liegt der Entscheidung für oder gegen eine Klasse selbstverständlich keine Inhaltsanalyse, wie sie ein menschlicher Katalogisierer durchführen würde, zugrunde. Stattdessen werden statistische Methoden, computerlinguistische Methoden oder solche des maschinellen Lernens angewendet. Die Analyse beruht dementsprechend nicht auf einem Verständnis des Inhalts, sondern auf der Erkennung von Mustern und Zusammenhängen.[34]

Das Ziel automatischer Klassifizierungsverfahren ist „die Approximation einer unbekannten Zielfunktion („target function"), die beschreibt, wie die Dokumente klassifiziert werden sollten. Dies geschieht mittels einer weiteren Funktion, die üblicherweise als *Klassifikator* („classifier") bezeichnet wird, und zwar so, dass diese beiden Modelle so gut wie möglich übereinstimmen."[35] Der Klassifikator steht also für das zur Klassifizierung verwendete Verfahren: „A classifier inputs a document and outputs a class."[36]

Prinzipiell können zwei Sichtweisen bei der automatischen Klassifizierung eingenommen werden:

a) Zu einem vorgegebenen Dokument werden eine oder mehrere passende Klassen ermittelt, denen das Dokument zugeordnet werden kann. In diesem Fall spricht man von *Dokumentenzentrierung*.

[33] Vgl. Oberhauser (2005), S. 20
[34] Vgl. Nohr (2005), S. 34
[35] Oberhauser (2005), S. 19
[36] Chakrabarti u.a. (1998), S. 167

b) Alternativ kann nach allen Dokumenten gesucht werden, die einer Klasse zugeordnet werden können. Dann spricht man von *Klassenzentrierung*.[37]

3.2 Entwicklung

Die ersten Versuche, Dokumente auf automatische Weise zu klassifizieren, fanden in den 1960er-Jahren statt.[38] Maron (1961) und Borko u.a. (1963) beschäftigten sich mit der Zuordnung von Dokumenten zu selbsterstellten Klassen, welche anhand von Termen aus den Abstracts gebildet wurden. Diese Klassen waren noch recht breit und allgemein definiert. Spätere Versuche, solche vordefinierten Klassen zu benutzen, finden sich z.B. bei Hoyle (1973) und Kar u.a. (1978). Der Großteil der Arbeiten dieser Zeit untersuchte jedoch das Clustering von Dokumenten, betrachtete also die Gruppierung ähnlicher Dokumente und gab keine Klassen vor. Methodisch war der Blick auf die Klassifizierungsaufgabe zunächst von der Theorie des Information Retrieval geprägt, sodass vorrangig dessen Methoden zur Anwendung kamen. Ein neu zu klassifizierendes Dokument wurde als Anfrage an das IR-System verstanden. Diese Anfrage wurde durch einen Vektor aus Termgewichten repräsentiert. In einer Datenbank waren die die einzelnen bekannten Klassen repräsentierenden Klassenvektoren (Zentroiden) gespeichert, mit denen die Anfrage nun verglichen wurde. Daraufhin konnte eine an der Ähnlichkeit von Anfrage und Zentroid orientierte Rangliste erstellt werden.[39]

Garland (1983) ermittelte Cluster anhand der Buchtitel und Schlagwörter und fand einen starken Zusammenhang zwischen den automatisch generierten Clustern und den intellektuell vergebenen Klassen der *Klassifikation der Library of Congress* (LCC). Enser (1985) verglich die Retrieval-Qualität von Klassen, die durch Clustering-Verfahren gewonnen wurden, und den durch menschliche Katalogisierer vergebenen Klassen. Er kam zu dem Schluss, dass automatisch generierte Klassen die relevante Literatur effektiver auffinden konnten.

In den 1980er-Jahren herrschte methodisch dann ein Ansatz vor, der die Implementierung von Expertensystemen vorsah. Manuelle, von Experten erstellte Regeln entschieden über die Zuordnung eines Dokumentes zu einer Klasse. Gut zehn Jahre später entwickelten sich Verfahren, die durch maschinelles Lernen eine Klassifizierung der vorhandenen Dokumente erreichen wollten.[40] Dies ist bis heute der vorherrschende Ansatz[41], daher wird dieser bei der

[37] Vgl. Oberhauser (2005), S. 19
[38] Der folgende Überblick orientiert sich an den Ausführungen bei Larson (1992), S. 131
[39] Vgl. Salton u.a. (1987), S. 145f., zit. nach Oberhauser (2005), S. 17f.
[40] Vgl. Oberhauser (2005), S. 17f.
[41] Vgl. ebd., S. 22

im weiteren Verlauf vorgenommenen Erläuterung automatischer Klassifizierung zugrunde gelegt.

3.3 Anwendungsgebiete

Sebastiani nennt fünf Anwendungsgebiete der automatischen Klassifizierung[42]:

Automatisches Indexieren für Boolesche Textretrievalsysteme: Hierbei werden den Termen eines kontrollierten Vokabulars passende Dokumente zugeordnet.

Dokumentenorganisation: Hierunter werden weitere Zuordnungen zur Strukturierung von Dokumentenbeständen, wie z.B. die thematische Gruppierung von Zeitungsanzeigen, verstanden.

Thematisches Textfiltern: Wenn regelmäßig neu eintreffende Dokumente (z.B. E-Mails) automatisch einer Kategorie oder einem Thema zugeordnet werden, spricht man vom Textfiltern.

Disambiguierung der Bedeutung von Homonymen bzw. Polysemen: Um gleichlautende oder -geschriebene Wörter ihrer richtigen Bedeutung im Kontext zuzuordnen, kann der Begriff als Dokument und der Kontext als Klasse betrachtet werden.

Hierarchisches Klassifizieren von elektronischen (v.a. Web-)Dokumenten: In dieses Anwendungsgebiet fallen die im zweiten Teil dieser Arbeit durchgeführten Untersuchungen.

3.4 Alternativen

Neben dem automatischen Klassifizieren gibt es weitere Techniken, welche die Anreicherung von Katalogdatensätzen mit Notationen einer Systematik zur Folge haben. Aufgrund ihrer möglichen Relevanz für die Beantwortung der Fragestellung der vorliegenden Arbeit wird im Folgenden auf diese eingegangen.[43]

a) *Fremddatenübernahme*: Durch Übernahme der Metadaten anderer Kataloge kann auf schnellem und einfachem Weg der eigene Katalog angereichert werden. Ein Beispiel für die Verwendung fremder Erschließungsdaten ist die Übernahme von Notationen der *Dewey-Dezimal-Klassifikation* (DDC) und der LCC aus z.B. dem World-

[42] Vgl. Sebastiani (2002), S. 5-7
[43] Vgl. Oberhauser (2005), S. 105

Cat in Katalogisate des *Gemeinsamen Verbundkatalogs* (GVK). Auch durch die Verbundkatalogisierung profitieren die teilnehmenden Bibliotheken von den bereits durch andere Katalogisierer angelegten Metadaten.

b) *Kooperation*: Bibliotheken können auch über (kommerzielle) Beziehungen zu anderen Dienstleistern an Erschließungsdaten gelangen. So können z.B. über einen ISBN-Abgleich gewünschte Metadaten aus einem Datenbestand in den eigenen Katalog importiert werden.

c) *Konkordanzen zwischen Klassifikationssystemen*: Möchte man den Anteil der mit einer bestimmten Systematik erschlossenen Bestände erhöhen und hat gleichzeitig bereits viele dieser Dokumente mit anderen Systematiken erschlossen, bietet sich auch die Verwendung von Konkordanzen an.[44]

3.5 Aufbau eines automatischen Klassifizierungssystems

Im Prozess des automatischen Klassifizierens werden zwei Hauptphasen unterschieden:

a) Die *Trainingsphase*: Hier wird aus vorhandenen Dokumenten Wissen extrahiert und der Klassifikator erstellt. Grundlage sind *Trainingsdokumente* („training set"), die zwangsläufig bereits mit dem im späteren Verlauf der automatischen Klassifizierung zu ermittelnden Zielattribut erschlossen sein müssen. „Ein zu einer bestimmten Klasse zählendes Dokument gilt als positives Beispiel, ein nicht zu dieser Klasse zählendes als negatives Beispiel; manche Algorithmen nutzen nur positive Beispiele, andere können von beiden Typen Gebrauch machen."[45]

b) Die *Klassifizierungsphase*: Hier wird unerschlossenen Dokumenten durch den Klassifikator nun ein Wert des Zielattributs zugewiesen. Dieser wird durch Vergleich mit den in der Trainingsphase erstellten Klassenprofilen ermittelt.[46]

[44] Es existieren einige Projekte und Veröffentlichungen zu diesem Thema: Beschreibungen von Konkordanzprojekten im Österreichischen Bibliothekenverbund finden sich im Vortrag von Plößnig u.a. auf dem RVK-Anwendertreffen 2014. Balakrishnan (2015) von der Verbundzentrale des GBV stellte im September 2015 auf der „European Conference on Data Analysis" die Arbeiten an „Cocoda (Colibri Concordance Database) – A mapping tool for library classification schemes" vor. Erkenntnisse aus einem Konkordanz-Projekt zwischen den medizinischen Klassen der DDC und der RVK von 2011 liegen ebenfalls vor (vgl. Balakrishnan (2011) und (2013)). Auf die Erstellung und Nutzung von Konkordanzen kann im Rahmen dieser Arbeit nicht weiter eingegangen werden. Es sei aber erwähnt, dass sie einen interessanten Ansatz zur Weiterentwicklung und Ergänzung maschineller Lernverfahren zur Klassifizierung von Dokumenten bilden können. Inwiefern beides zusammenhängt, zeigen die Untersuchungen im zweiten Teil der vorliegenden Arbeit. Zur Einführung in das Thema bieten sich der Aufsatz „Die Konkordanz von Klassifikationen – hat sie eine Chance?" von Hermes (1996) sowie der darin zitierte Artikel von Nöther (1994a) und (1994b) an.

[45] Oberhauser (2005), S. 22

[46] Für die Beschreibung der beiden Phasen vgl. ebd., S. 17f.

Um die Güte des erstellten Klassifikators beurteilen und vergleichen zu können, werden in der Trainingsphase mind. zwei unterschiedliche Dokumentenmengen verwendet: Neben den *Trainingsdokumenten* zur Erstellung des Klassifikators werden *Testdokumente* benötigt, mit deren Hilfe die Qualität beurteilt werden kann. Diese Dokumente müssen bereits intellektuell klassifiziert worden sein. Nach Erstellung des Klassifikators wird dann jedes Testdokument automatisch klassifiziert und dieses Ergebnis mit der in der intellektuellen Klassifizierung vergebenen Klasse verglichen. So kann z.B. beurteilt werden, wie viel Prozent der Dokumente richtig – d.h. übereinstimmend mit der intellektuellen Klassifizierung – klassifiziert wurden. Falls während der Trainingsphase unterschiedliche Versionen oder Parameter des Klassifikators miteinander verglichen werden sollen, ist die Entnahme einer Menge von *Validierungsdokumenten* aus der Menge der Trainingsdokumente nötig. Mit diesen können dann unterschiedliche Varianten des Klassifikators bezüglich ihrer Leistungsfähigkeit gegenübergestellt werden.

Die Testdokumente sind i.d.R. weniger zahlreich als die Trainingsdokumente. Selbstverständlich gilt, dass in der Klassifizierungsphase nur diejenigen Klassen vergeben werden können, die in der Trainingsphase gelernt werden konnten. Daher sollte bei der Zusammenstellung der Trainingsdokumente berücksichtigt werden, dass aus jeder Klasse Vertreter enthalten sind.[47] Weiterhin dürfen die Testdokumente nicht gleichzeitig Teil der Trainingsdokumente sein, um die Evaluierung des Klassifikators nicht zu verfälschen. Nach der Evaluierung wird der Klassifikator vor seinem produktiven Einsatz normalerweise ein weiteres Mal auf Basis der um die Testdokumente vergrößerten Trainingsdokumente trainiert. So wird eine möglichst große Anzahl von Trainingsdokumenten erreicht.[48]

3.6 Forschungsstand

Anschließend an die Begriffsklärungen und Erläuterungen der vorangegangenen Abschnitte wird nun der aktuelle Forschungsstand auf dem Gebiet der automatischen Klassifizierung betrachtet.

Kompakte Zusammenfassungen, die sich explizit dem Themenbereich der automatischen Klassifizierung von Bibliotheksbeständen annehmen, sind selten. Eine deutliche Mehrheit der untersuchten Literatur bezieht sich auf die Klassifizierung von elektronischen Dokumenten bzw. Webressourcen. Diese ermöglichen zumeist die Verwendung von Volltexten oder

[47] Dies stellt eine Schwierigkeit bei Datenbeständen mit vielen schwach belegten Klassen dar, wie im zweiten Teil dieser Arbeit noch zu sehen sein wird.
[48] Vgl. Oberhauser (2005), S. 22f.

zumindest Abstracts. Forschungen, die sich ausschließlich auf die „klassischen“ Metadaten eines Bibliothekskatalogs stützen, sind kaum vorhanden.

Nohr (2005) bietet eine gute Einführung in das Themengebiet der automatischen Indexierung und beschreibt die wesentlichen Methoden, Indexierungsverfahren und Evaluationsmöglichkeiten. Dies geschieht allerdings mehr aus wirtschaftsinformatischer als aus bibliothekarischer Sicht. Da die Grundlagen jedoch fachübergreifend sind, bietet sich dieses Lehrbuch auch für Informationswissenschaftler als Einstieg an. Im selben Jahr veröffentlichte Oberhauser eine ausführliche Übersicht, die sich explizit mit dem Thema der automatischen Klassifizierung auseinandersetzt und andere Indexierungsverfahren nur am Rande erwähnt. In seiner Beschreibung der Methodik des automatischen Klassifizierens beruft er sich dabei weitestgehend auf einen Artikel von Sebastiani (2002) mit dem Titel „Machine Learning in Automated Text Categorization“. Dieser kann aufgrund seines Umfangs und seiner häufigen Zitierung als wichtigste Quelle zur Einführung in die Verwendung von maschinellem Lernen bei der automatischen Klassifizierung betrachtet werden. Neben der Methodik des automatischen Klassifizierens beschreibt Oberhauser mehrere Projekte aus der Zeit um die Jahrtausendwende, die sich jedoch größtenteils mit der Klassifikation von elektronischen (Web-) Ressourcen beschäftigen. In Kapitel 7 (S. 99ff.) betrachtet Oberhauser Projekte, die sich mit der Klassifizierung von Büchern beschäftigt haben. Er konstatiert, dass nur sehr wenig Literatur zur automatischen Klassifizierung von Büchern existiert und aus dem Bibliotheksbereich hierzu „keine signifikanten Studien oder Anwendungen“[49] vorliegen.[50]

Einzig die Untersuchung von Larson (1992) ist besonders hervorzuheben. Diese beschäftigt sich ausführlich mit der automatischen Zuordnung von Klassen der LCC zu MARC-Datensätzen. Da dieser Artikel als wesentliche Grundlage der im weiteren Verlauf dieser Arbeit durchgeführten Untersuchungen dient, werden die dort verwendeten Methoden und Ergebnisse im Folgenden dargestellt:

Die Untersuchung basierte auf rund 30.000 MARC-Datensätzen aus der bibliothekswissenschaftlichen Fachbibliothek der University of California. Diese waren durch die LCC erschlossen, wobei ca. 92 % der Datensätze eine Klasse aus der Hauptklasse Z aufwiesen. Insgesamt waren 5.765 unterschiedliche Klassen vorhanden. Für jede dieser Klassen wurden

[49] Oberhauser (2005), S. 139
[50] Im Gegensatz dazu finden sich reichlich Untersuchungen, die Experimente zur automatischen Vergabe von Schlagwörtern oder Deskriptoren eines Thesaurus durchführen. Siehe hierfür z.B. Mittelbach u.a. (2006).

„classification cluster“ gebildet, welche die in der Klasse vorhandenen Dokumente repräsentierten. Dies geschah in der Form von Vektoren von Attributgewichten. Die den Klassen zugeordneten Begriffe aus den Tafeln der LCC wurden hingegen nicht verwendet.

Anschließend wurden 283 neue, bereits intellektuell erschlossene Datensätze anhand der gebildeten Cluster klassifiziert. Die berechneten Klassen wurden nach ihrem Grad der Übereinstimmung gereiht. Durch die bereits vorhandene intellektuelle Erschließung konnte das Klassifizierungsergebnis mit der tatsächlich vergebenen Klasse verglichen werden. Larson verwendete vier verschiedene Methoden der Berechnung von Termgewichten, fünf Varianten der Attributauswahl, zwei Stemming-Methoden und einen Ansatz zur Normalisierung von Phrasen, sodass insgesamt 60 Klassifikationsverfahren getestet und ihre Ergebnisse verglichen wurden. Die einzelne Darstellung dieser Verfahren würde hier jedoch zu weit führen.

Mit der besten Kombination der o.g. Methoden wurde eine Genauigkeit der Klassenzuordnung von 46,6 % erreicht. Dies bedeutet, dass von dem automatischen System in nahezu jedem zweiten Fall die auch intellektuell vergebene Klasse ermittelt wurde. Da eine Rangfolge der den Testdokumenten ähnlichsten „classification cluster“ gebildet wurde, ließ sich darüber hinaus feststellen, dass in 74,4 % der Fälle die korrekte Klasse unter den zehn besten Ergebnissen vorzufinden war. Larson schloss, dass eine automatische Vergabe von LCC-Notationen nicht möglich sei, eine semi-automatische Sacherschließung jedoch zufriedenstellende Resultate bringen würde.[51]

Im Folgenden eine kurze Darstellung weiterer von Oberhauser genannten Untersuchungen zur Klassifizierung von Büchern:

Cheng u.a. (1995): Hier wurde unter Verwendung von Titel und Kapitelüberschriften der Bücher klassifiziert. Die betrachtete Klassifikation war die DDC. Es konnten zwischen 85 und 90 % an korrekten Zuordnungen erreicht werden. Die Autoren verwendeten jedoch nur eine kleine Kollektion.

Ishida (1998): Hier fand eine automatische Zuordnung zu Klassen der japanischen *Nippon Decimal Classification* (NDC) statt. Der Studie lag eine Kollektion von 1.000 Büchern zugrunde, verschiedene Extraktions- und Gewichtungsmethoden wurden getestet. Das beste Ergebnis bestand in 55,9 % korrekter Zuordnungen.[52]

[51] Vgl. für die Zusammenfassung der Untersuchung von Larson Oberhauser (2005), S. 99-101
[52] Die Angaben entstammen dem Abstract, da die Studie nur in japanischer Sprache vorliegt.

Andere Forschungen sind nur schlecht dokumentiert oder es liegen nur Abstracts vor, sodass bis zum Jahr 2005 tatsächlich nur wenig relevante Literatur zum automatischen Klassifizieren von Büchern vorlag. In den vergangenen zehn Jahren erschienen ebenfalls nur wenige aussagekräftige Veröffentlichungen zu diesem Thema.

Der von Pong u.a. (2008) veröffentlichte Aufsatz mit dem Titel „A comparative study of two automatic document classification methods in a library setting" beschäftigt sich mit den Schwierigkeiten, die bibliothekarische Datensätze im Allgemeinen und die LCC im Speziellen für die automatische Generierung von Notationen aufwerfen. Sie vergleichen die beiden Algorithmen *k-nearest-neighbours* (KNN) und *Naïve Bayes* und stellen ein selbstentwickeltes „automatic document classification system, WADCS"[53] vor. Zur Verbesserung der Klassifizierungsleistungen nutzen sie hierfür eine von den Autoren aufbereitete Version der LCC. Sie schließen mit der Aussage, dass KNN sich zur Unterstützung des konventionellen Erschließungsprozesses besser eignet als Naïve Bayes.

Wang (2009) untersucht die Vergabe von Klassen der DDC mithilfe von Verfahren des überwachten Lernens. Einer gründlichen Analyse der Verteilung der Trainingsdokumente über die DDC folgt die Konzeptionierung einer veränderten DDC-Struktur, um der Klassifikation inhärente Probleme zu verringern. Damit eine akzeptable Güte der Klassifizierungsleistung erreicht wird, wird ein semi-automatisches System vorgeschlagen, das bei maximal drei Benutzer-Interaktionen eine korrekte Klassifizierungsrate von 90 % erreicht.

In „A Review of Cataloging and Classification Literature 2011-12" von Martin u.a. (2014) wird nur eine Veröffentlichung erwähnt, die sich mit der automatischen Klassifizierung beschäftigt: Joorabchi u.a. (2011) betrachten die zwischen Dokumenten bestehenden Verlinkungen in Form von Zitationen und versuchen, darauf aufbauend eine Zuordnung von elektronischer Literatur zu den Klassen der DDC zu erreichen.

Die Aktualität des Themas wird auch dadurch deutlich, dass in den Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft nahezu in jedem der letzten Jahre thematisch verwandte Veröffentlichungen stattfanden; z.B. finden sich Projekte zur Anreicherung von klassischen Bibliotheksanwendungen mit automatischen Erschließungsdaten.[54]

Literatur, die sich mit der automatischen Klassifizierung von Dokumenten zu Klassen der *Klassifikation der National Library of Medicine* (NLMC) beschäftigt, wurde während der

[53] Pong u.a. (2008), S. 227
[54] Vgl. Schneider (2008)

für die vorliegende Arbeit stattfindenden Literaturrecherche nicht gefunden. Vielmehr bestätigt sich auch bei dieser Klassifikation der Gesamteindruck: es existieren einige Veröffentlichungen zur automatischen Indexierung[55], also in diesem Fall der Zuordnung von Termen des von der NLM verwendeten Thesaurus *Medical Subject Headings* (MeSH), aber keine explizit der Klassifizierung gewidmeten Arbeiten.

Zusammenfassend zeigt sich, dass durchaus diverse Ansätze zur Erforschung der automatischen Klassifizierung von Bibliotheksbeständen bestehen. Allerdings sind diese häufig nicht miteinander vergleichbar, da sie z.B. verschiedene Klassifikationssysteme nutzen. Und selbst bei ähnlichen Voraussetzungen unterscheiden sich die Forschungen teilweise stark in Methodik, Datenstruktur, -verarbeitung und Dokumentation. Dass nur wenig relevante Literatur existiert, wird auch dadurch deutlich, dass viele der o.g. Autoren ihre Untersuchungen als bislang einzigartig in diesem Bereich bezeichnen.

4 Klassifikationen

Klassifikationen im Bibliothekskontext haben die Aufgabe, thematisch ähnliche Literatur in Klassen zusammenzufassen und dadurch von thematisch unähnlichen Ressourcen (in anderen Klassen) abzugrenzen. Dies geschieht zur Unterstützung des Nutzers bei seiner Suche nach relevanter Literatur.[56] Neben dieser abstrakten Funktion sorgen Klassifikationen auch dafür, die Bestände einer Bibliothek physisch zu ordnen, wenn sie als Aufstellungssystematiken zur Organisation der Bestände auf den Regalen der Bibliothek dienen. Hierdurch wird das abstrakte Klassenkonzept sichtbar gemacht: „Library classification can show the distance between separate subjects by the distance between books on those subjects on the library shelves. The library becomes a physical embodiment of a knowledge structure.“[57]

Um diese Ordnung sicherzustellen und jedes Buch individuell identifizierbar zu machen, muss ein Klassifikationssystem die Möglichkeit besitzen, die einzelnen Klassen (und innerhalb dieser das einzelne Buch) eindeutig zu kennzeichnen. Damit diese Kennzeichnung möglichst leicht verständlich ist, werden hierfür – i.d.R. alphanumerische – Codes bzw. Symbole verwendet. Diese werden auch als *Notationen* bezeichnet.[58] Da sich sowohl Zahlen als auch Buchstaben in einer definierten Weise sortieren lassen, ermöglichen Notationen die schnelle Einordnung eines Themas in das Gesamtkonzept einer Klassifikation: „Notation is

[55] z.B. Vasuki u.a. (2010)
[56] Vgl. Batley (2005), S. 3
[57] Ebd.
[58] Vgl. ebd., S. 4

the group of symbols, technically applied, which as a code represent the subjects contained in the schedules of a classification scheme in order that these subjects will be filed at the correct point in a physical sequence of subjects."[59] Eine Übersicht über die Klassenbezeichnungen einer Klassifikation und ihrer dazugehörigen Notationen wird als Tafel (engl. „schedule") bezeichnet.[60]

4.1 Arten von Klassifikationen

Je nach Struktur unterscheidet man zwei Arten von Klassifikationen: *enumerative Klassifikationen* und *Facettenklassifikationen*.[61] Weitere gebräuchliche Bezeichnungen für erstere sind z.B. „hierarchische Klassifikation", „präkombinierte Klassifikation" oder „analytische Klassifikation".[62]

Enumerative Klassifikationen zeichnen sich dadurch aus, dass Elemente einer Klasse gleichzeitig auch Elemente der übergeordneten Klasse(n) sind. Zu jeder Klasse gehört nur eine Oberklasse und die einzelnen Klassen sind stark ausdifferenziert, was häufig zu langen und detaillierten Klassenbezeichnungen führt.[63] Dies stellt sicher, dass möglichst viele Aspekte eines Themas eingeordnet werden können. Durch die Auflistung aller vorhandenen Klassen werden gleichzeitig alle möglichen Notationen dargestellt. Von den im folgenden Abschnitt betrachteten Klassifikationen gilt die LCC als „most fully enumerative".[64]

Facettenklassifikationen betrachten die zu klassifizierenden Objekte aus verschiedenen Perspektiven und zerlegen das betrachtete Thema in einzelne Aspekte. Diese werden beim Klassifizieren wieder zusammengesetzt, sodass dadurch komplexe Sachverhalte ausgedrückt werden können. Dementsprechend sind Facettenklassifikationen flexibler in ihrer Handhabung und besitzen im Gegensatz zur enumerativen Klassifikation die Verknüpfung als ein wesentliches Merkmal.[65] Statt der Aufzählung sämtlicher Aspekte eines Themas werden also die hauptsächlichen Eigenschaften eines zu klassifizierenden Objektes abgebildet. Dies entspricht den Erkenntnissen moderner Klassifikationstheorie.[66]

[59] Marcella u.a. (1994), zit. nach Batley (2005), S. 25
[60] Vgl. Batley (2005), S. 19
[61] Die entsprechenden englischen Bezeichnungen lauten „enumerative" und „faceted classification"
[62] Vgl. Bertram (2005), S. 167, Fußnote 2
[63] Vgl. ebd., S. 167
[64] Chan (2007), S. 312
[65] Vgl. Bertram (2005), S. 173-175
[66] Vgl. Chan (2007), S. 312

Die obigen Ausführungen stellen indes nur idealtypische Ausprägungen der beiden Typen dar. In den meisten in der Praxis verwendeten Klassifikationen vermischen sich Prinzipien sowohl der enumerativen als auch der Facettenklassifikation.[67]

4.2 Klassifikationen im untersuchten Datenbestand

Die im Rahmen dieser Arbeit untersuchten Datensätze weisen Zuordnungen zu unterschiedlichen Klassifikationen auf. Daher werden diese im Folgenden jeweils genauer dargestellt. Neben der Geschichte, grundsätzlichen Prinzipien und der allgemeinen Zusammensetzung der Notationen wird die Art und Weise der Klassifizierung medizinischer Literatur beschrieben.[68]

4.2.1 Klassifikation der Library of Congress

Geschichte und Prinzipien

Die *Klassifikation der Library of Congress* (LCC) ist eine analytische Klassifikation. Sie wurde um das Jahr 1900 entwickelt, um die Bestände der Library of Congress zu organisieren. Im Laufe des 20. Jahrhunderts wurde sie von weiteren, vorwiegend großen wissenschaftlichen Bibliotheken der USA zur Bestandserschließung übernommen. Ihre Verwendung in zahlreichen Bibliotheken außerhalb der USA macht die LCC nach der DDC zur weltweit zweitmeistgenutzten Klassifikation.[69]

Da die LCC nicht mit dem Anspruch erstellt wurde, von anderen Bibliotheken übernommen zu werden, ist sie an den Erfordernissen der Library of Congress orientiert und basiert nicht durchgehend auf Erkenntnissen zur Theorie der Klassifikation bzw. der Wissensstrukturierung. Ihre Struktur ist daher nicht rein logisch, sondern pragmatisch.[70] Die LCC entspricht weniger einer Darstellung von Wissen und dessen Strukturen als vielmehr einer detaillierten Auflistung von Themen. Ihre Beliebtheit resultiert wahrscheinlich u.a. aus der Tatsache, dass sie dank dieser umfangreichen Auflistung auf das individuelle Zusammensetzen von Notationen durch den Katalogisierer verzichtet. In neueren Ausgaben der LCC wird jedoch ver-

[67] Erläuterungen zu den Vor- und Nachteilen der beiden Klassifikationstypen finden sich u.a. bei Batley (2005), S. 4-9
[68] Eine detaillierte Beschreibung der jeweiligen Vor- und Nachteile einzelner Klassifikationen soll an dieser Stelle nicht geschehen. Diese findet sich z.B. bei Chan (1999). Auf einige Unreinheiten innerhalb der LCC macht Batley (2005), S. 69 aufmerksam.
[69] Vgl. Library of Congress (2014)
[70] Vgl. Batley (2005), S. 60

mehrt die Anwendung von Tabellen (also Elementen einer Facettenklassifikation) eingeführt. Hierdurch lassen sich die vorhandenen Klassen um einzelne wichtige Aspekte (z.B. geographische oder zeitliche Angaben) erweitern.[71]

Die LCC enthält eine Verzerrung in Richtung amerikanischer Themen bzw. westlicher Sichtweisen, weswegen sie außerhalb der USA nicht so verbreitet ist wie z.B. die DDC.[72] Ein wesentlicher Nachteil der LCC ist, dass die erfassten Inhalte auf den Erwerbungen der Library of Congress basieren, sodass neue Klassen nur nach der Publikation und Akquise entsprechender Literatur ergänzt werden.[73]

Aufbau der Notationen

Auf oberster Ebene besteht die LCC aus 21 Basisklassen, die jeweils durch einen Buchstaben des lateinischen Alphabets repräsentiert werden. Diese können auf der zweiten Ebene durch Hinzufügen von ein bis zwei weiteren Buchstaben weiter untergliedert werden. Auf der dritten Hierarchieebene sind die jeweiligen Bereiche in einer lose hierarchischen Ordnung von allgemeinen zu spezifischeren Themen sortiert. Oft werden die einzelnen Themen auch nach bestimmten geographischen Einheiten, bibliographischen Erscheinungsformen oder Zeiträumen gegliedert.[74]

Jedem Thema ist eine ein- bis vierstellige Nummer oder ein Nummernbereich zugeordnet. Gelegentlich werden diese Nummern durch Anfügen eines Dezimalpunktes weiter untergliedert. Manche Themen sind nicht hierarchisch, sondern alphabetisch sortiert. In diesem Fall werden sie durch Kombination eines Buchstabens und einer Nummer repräsentiert.[75] Falls vorhanden, kann zur weiteren Spezifizierung des Inhalts eine Zahlen-Buchstaben-Kombination (die *Cutter-Nummer*) hinzugefügt werden. Aus all diesen Elementen entsteht dann die sog. „class number".[76]

Beziehungen zwischen den Klassen werden nicht durch ihre „class number", sondern durch die Einrückung der Klassenbeschreibung innerhalb der Tafeln der Klassifikation verdeutlicht.[77] Die Klasse „QH – Ecology" enthält z.B. die Unterklassen „QH541 – General works, treatises, and textbooks" und „QH541.13 – Popular works". In diesem Fall kann – anders als z.B. bei der DDC – nicht anhand der Notation auf die Hierarchie der Klassen geschlossen

[71] Vgl. ebd., S. 70
[72] Vgl. ebd., S. 60
[73] Vgl. ebd., S. 70
[74] Vgl. Library of Congress (2014)
[75] Vgl. ebd.
[76] Vgl. Batley (2005), S. 74
[77] Vgl. Library of Congress (2014)

werden. Ein Blick in die Klassenbeschreibung zeigt, dass die Klassen QH541 und QH541.13 auf derselben Ebene angeordnet sind (Abb. 1). Der Grund für die dezimale Unterteilung liegt allein darin, dass die Klasse QH541.13 nicht von Anfang an in der LCC vorhanden war und später an einer passenden Stelle hinzugefügt werden musste.[78]

541	General works, treatises, and textbooks
541.13	Popular works
541.14	Juvenile works
541.142	Handbooks, tables, formulas, etc.
541.145	Addresses, essays, lectures
541.15.A-Z	Special aspects of the subject as a whole, A-Z
541.15.A9	Autoradiographic techniques

Abb. 1: Hierarchische Beziehungen in der LCC [79]

Da die LCC als Aufstellungssystematik für die Library of Congress konzipiert wurde, kann mit ihrer Hilfe eine Ressource eindeutig identifiziert werden. Dies geschieht durch die „call number", welche aus der Kombination von „class number" und „book number" besteht. Die „book number" setzt sich aus der Cutter-Nummer für den Haupteintrag (entweder Autor oder Titel) und dem Jahr der Veröffentlichung zusammen (Abb. 2).[80]

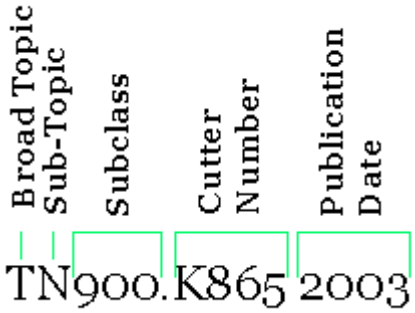

Abb. 2: Beispiel einer call number aus der LCC [81]

Medizinische Ressourcen in der LCC

Die LCC verwendet für den Bereich Medizin den Buchstaben R, der in insgesamt 17 Hauptklassen untergliedert wird.[82]

[78] Das Beispiel entstammt Robare u.a. (o.J.)
[79] Screenshot aus der online verfügbaren PDF-Datei der Klasse Q, https://www.loc.gov/aba/publications/FreeLCC/Q-text.pdf, zuletzt geprüft am 24.08.2016.
[80] Vgl. Batley (2005), S. 74
[81] http://www.zackgrossbart.com/hackito/wp-content/uploads/2007/11/lcc.gif, zuletzt geprüft am 24.08.2016.
[82] Vgl. Library of Congress (o.J.)

4.2.2 Dewey-Dezimal-Klassifikation

Geschichte und Prinzipien

Die *Dewey-Dezimal-Klassifikation* (DDC) zählt zu den Universalklassifikationen und ist die weltweit verbreitetste bibliothekarische Klassifikation. Sie wurde vom Bibliothekar Melvil Dewey entwickelt und das erste Mal im Jahr 1876 veröffentlicht.[83] Die englische Version existiert mittlerweile in der 23. Ausgabe aus dem Jahr 2011. Die Deutsche Nationalbibliothek verwendet zur Sacherschließung seit 2006 die Übersetzung der 22. Ausgabe aus dem Jahr 2003.[84] Durch die breite Abdeckung vieler Themenbereiche eignet sich die DDC gut für die Klassifikation der Bestände von (kleineren) Universalbibliotheken, wie z.B. Schul- oder öffentlichen Bibliotheken. Bei der Erschließung speziellerer, fachspezifischer Kollektionen stößt sie jedoch schnell an ihre Grenzen.[85] Die Struktur der DDC ist in manchen Bereichen allerdings auch heute noch vom kulturellen, westlichen Hintergrund ihrer Entwickler geprägt. So finden sich beispielsweise in der Klasse „200 – Religion“[86] unter den Klassen 200-280 christliche Themen, sämtliche anderen Religionen sind unter der Klasse 290 zusammengefasst.[87]

Aufbau der Notationen

Als Dezimalklassifikation unterteilt die DDC ihre Klassen in jeweils 10 Unterklassen. Somit gibt es auf oberster Ebene 10 Klassen, auf zweiter Ebene 100 und auf der dritten Ebene insgesamt 1000 Klassen. Weitere Unterteilungen können dann durch das Anfügen eines Dezimalpunktes und weiterer Ziffern erfolgen. Anhand von vorgegebenen Regeln können zusätzliche Unterteilungen konstruiert werden, z.B. für geographische Gebiete oder Publikationsformen. Somit listet die DDC (im Gegensatz zur LCC) nicht sämtliche möglichen Themen auf, sondern überlässt dem Katalogisierer teilweise die Konstruktion der Notationen.[88]

Ein Werk erhält immer eine mindestens dreistellige Notation. So steht „600“ beispielsweise für den Bereich „Technik“, 610 für die Unterklasse „Medizin und Gesundheit“ und 611 für deren Unterklasse „Menschliche Anatomie, Zytologie, Histologie“ (Abb. 3). Somit kann anhand der Notation (anders als bei der LCC) die Hierarchie der Klassen abgebildet werden.

---

[83] Vgl. OCLC (o.J.)
[84] Vgl. Deutsche Nationalbibliothek (o.J.)
[85] Vgl. Batley (2005), S. 27f.
[86] Die Bezeichnungen der Klassen richten sich nach der deutschen Übersetzung der DDC, ermittelt durch WebDeweySearch, online verfügbar unter http://deweysearchde.pansoft.de/webdeweysearch/mainClasses.html, zuletzt geprüft am 24.08.2016.
[87] Vgl. Batley (2005), S. 28
[88] Vgl. ebd., S. 34

Eine Klasse ist immer in derjenigen Klasse enthalten, die eine Ziffer kürzer ist.[89] Nach einem Dezimalpunkt (welcher nur der Übersichtlichkeit dient) können weitere Ziffern zur Ausdifferenzierung des Themengebiets angehängt werden, wobei jede DDC-Nummer mit mehr als drei Stellen nicht mit einer 0 enden darf.[90]

600	Technik, Medizin, angewandte Wissenschaften
610	Medizin & Gesundheit
611	**Menschliche Anatomie, Zytologie, Histologie** • Für pathologische Anatomie siehe Pathologie
611.001-611.009	Standardschlüssel
611.01	Anatomische Embryologie, Zytologie, Histologie
611.1-611.9	Makroskopische Anatomie

Abb. 3: Hierarchische Beziehungen in der DDC [91]

Wird die DDC als Aufstellungsnotation benutzt, geschieht die eindeutige Identifizierung einer Ressource durch eine „call number“, welche aus der Kombination von „class number“ und „book number“ besteht. Die „book number“ setzt sich aus der Cutter-Nummer für den Haupteintrag (entweder Autor oder Titel) und dem Jahr der Veröffentlichung zusammen.[92]

Medizinische Ressourcen in der DDC

Literatur aus dem Themenbereich Medizin wird in der Klasse „610 – Medizin und Gesundheit“ klassifiziert. Diese ist in 9 weitere Klassen unterteilt, wobei Klasse „619“ nicht vergeben ist (siehe Abb. 4). Klasse „617“ enthält beispielsweise „Chirurgie, Medizin nach Körperregion, Zahnmedizin, Augenheilkunde, Ohrenheilkunde und Audiologie“. Unter „617.001-617.5“ finden sich dann ausschließlich chirurgische Themen, die noch weiter unterteilt werden.

[89] Vgl. OCLC (o.J.)
[90] Vgl. Batley (2005), S. 42
[91] Screenshot aus WebDeweySearch, online verfügbar unter http://deweysearchde.pansoft.de/webdeweyse-arch/mainClasses.html, zuletzt geprüft am 24.08.2016.
[92] Vgl. Chan (2007), S. 362-372. Siehe auch S. 24 der vorliegenden Arbeit.

610	Medizin & Gesundheit
610	Medizin und Gesundheit
611	Menschliche Anatomie, Zytologie, Histologie
612	Humanphysiologie
613	Persönliche Gesundheit und Sicherheit
614	Rechtsmedizin; Inzidenz von Verletzungen, Wunden, Krankheiten; öffentliche Präventivmedizin
615	Pharmakologie und Therapeutik
616	Krankheiten
617	Chirurgie, Medizin nach Körperregion, Zahnmedizin, Augenheilkunde, Ohrenheilkunde, Audiologie
618	Andere Fachrichtungen der Medizin Gynäkologie und Geburtsmedizin

Abb. 4: Medizinische Hauptklassen der DDC [93]

4.2.3 Klassifikation der National Library of Medicine

Geschichte und Prinzipien

Im Gegensatz zu den bislang betrachteten Universalklassifikationen, die nahezu den ganzen Bereich menschlichen Wissens abzudecken versuchen, handelt es sich bei der Klassifikation der *National Library of Medicine* (im Folgenden mit NLMC abgekürzt) um eine Spezialklassifikation, welche die Literatur eines thematisch begrenzten Bereiches zu erfassen versucht. Ihre erste Ausgabe wurde im Jahr 1951 unter dem Titel *U.S. Army Medical Library Classification* veröffentlicht. Seit 2002 erscheint die NLMC ausschließlich als elektronische Version mit jährlichen Aktualisierungen.[94] Sie wird auch außerhalb der *National Library of Medicine* (NLM) von vielen medizinischen Bibliotheken verwendet. In Deutschland verwendet neben der Medizinischen Hochschule Hannover nur die Bibliothek des Universitätsklinikums Hamburg-Eppendorf die NLMC.

Bei der Konzeption der NLMC wurde als ein Kriterium festgelegt, dass ihre Klassenstruktur sich nahtlos in die LCC einfügen sollte. Daher wurde in Absprache mit der Library of Congress vereinbart, die in der LCC bis dato nicht verwendete Klasse W und den unbelegten

[93] Screenshot aus WebDeweySearch, online verfügbar unter http://deweysearchde.pansoft.de/webdeweyse-arch/mainClasses.html, zuletzt geprüft am 24.08.2016.
[94] Vgl. NLM (2016)

Bereich QS-QZ für medizinische Literatur zu nutzen. Die Klassen QS-QZ (aus dem Bereich „Science“ der LCC) werden von der NLM seither für Literatur aus dem Themenspektrum der präklinischen Fächer vergeben; Klasse W und ihre Unterklassen für medizinische Literatur. Diese Klassen werden auch zukünftig innerhalb der LCC nicht belegt werden. Hierdurch sind sowohl eine Vermeidung der Doppelbelegung dieser Buchstabenbereiche als auch die vollständige Integration der NLMC in die LCC sichergestellt. Um letzteres zu erreichen, folgt die NLMC auch im Aufbau ihrer Notationen den Strukturen der LCC. In medizinischen Bibliotheken klassifizierte Literatur, die nicht in die Bereiche der NLMC fällt, wird in den regulären Klassen der LCC klassifiziert. Ausgenommen sind hiervon die Klassen QM (Anatomie), QP (Physiologie), QR (Mikrobiologie) und R (Medizin), da sich diese mit den Klassen der NLM überschneiden.[95]

Aufbau der Notationen

Die Notationen der NLMC bestehen in der Regel aus ein bis zwei Buchstaben, denen bis zu drei arabische Ziffern folgen können. Manche Klassen sehen darüber hinaus eine weitere Unterteilung mittels Dezimalpunkten vor (Abb. 5). Auch mit Cutter-Nummern können weitere Spezifizierungen vorgenommen werden.[96] Die NLMC beinhaltet nur eine ergänzende Tabelle, um den Notationen weitere Aspekte hinzuzufügen: In der „Table G“ werden für geographische Einheiten Kombinationen aus Buchstaben und Zahlen vorgegeben, die bei Belegung bestimmter Klassen der Notation hinzugefügt werden (siehe z.B. die Klasse „WC 503.4“ in Abb. 5).[97]

Acquired Immunodeficiency Syndrome. HIV Infections

WC 503	Acquired immunodeficiency syndrome. HIV infections
WC 503.1	Diagnosis
	Classify works on AIDS serodiagnosis in QY 265.
WC 503.2	Therapy
WC 503.3	Etiology. Transmission
WC 503.4	Epidemiology (Table G)
WC 503.41	General coverage (Not Table G)
WC 503.5	Complications
WC 503.6	Prevention and control
WC 503.7	Psychosocial aspects

Abb. 5: Notationen der NLMC [98]

[95] Vgl. ebd.
[96] Vgl. Chan (2007), S. 414
[97] Vgl. ebd., S. 415f.
[98] Screenshot aus der Klasse WC der NLMC, online verfügbar unter https://wwwsvlt.nlm.nih.gov/class/docs/class_wc.html, zuletzt geprüft am 24.08.2016.

Abweichend vom grundlegenden Schema werden zur Klassifizierung von Werken des 19. Jahrhunderts Folgen aus drei Buchstaben verwendet.[99] Alte Bestände (d.h. Literatur, die vor 1801 gedruckt wurde) werden in den Klassen WZ 230-270 eingeordnet. Nachdrucke solcher Literatur befinden sich in WZ 290-292, moderne Untersuchungen solcher Werke unter WZ 294.[100] Die ersten Klassen der zweiten Hierarchieebene mit den Nummern 1-39 sind für bestimmte Publikationstypen (z.B. Handbücher, Wörterbücher, Gesetze, Atlanten) aus dem betreffenden Bereich vorgesehen.[101]

Insgesamt betrachtet ist die NLMC eine recht breite Klassifikation, die ihre vollen Ausdrucksmöglichkeiten erst im Zusammenspiel mit den MeSH erreicht.[102]

Verwendung der NLMC in der MHH

Die Klassifizierung innerhalb der MHH entspricht weitestgehend den Vorgehensweisen der NLM. Serien werden jedoch abweichend nicht unter z.B. W 1 klassifiziert, sondern erhalten lediglich die übergeordneten Buchstaben der passenden Klasse (in diesem Beispiel also die Klasse W).

4.2.4 Basisklassifikation

Geschichte und Prinzipien

Die *Basisklassifikation* (BK) wurde Ende der 1980er-Jahre vom niederländischen PICA-Verbund entwickelt und erschien 1992 erstmals in einer deutschen Übersetzung. Seit 1993 wird sie vom Gemeinsamen Bibliotheksverbund verwendet. Die momentan gültige Fassung ist die 3. Ausgabe von 2000, deren 4. Ergänzungslieferung 2011 erschien. Auf erster Ebene beinhaltet die BK 48 Klassen, die in fünf Gebieten zusammengefasst sind. Insgesamt enthält sie ca. 2.100 Klassen und ist somit eine recht grobe Klassifikation. Sie wird in der Praxis daher durch die Verwendung weiterer Sacherschließungselemente (wie z.B. Schlagwörter) ergänzt.[103]

Aufbau der Notationen

Die 48 Hauptklassen werden durch eine zweistellige Ziffernfolge identifiziert. Zur Repräsentation ihrer Unterklassen werden ein Dezimalpunkt und eine weitere zweistellige Ziffernfolge angehängt.

---

[99] Vgl. ebd., S. 414
[100] Vgl. ebd., S. 418
[101] Vgl. NLM (2016)
[102] Vgl. Chan (2007), S. 414
[103] Vgl. Gemeinsamer Bibliotheksverbund (2011), S. V-VII

Medizinische Ressourcen in der BK

Medizinische Literatur wird in der Hauptklasse 44 zusammengefasst. Insgesamt existieren hier 79 Unterklassen. Aus der Notation lässt sich jedoch nicht – anders als bei der DDC – durchgehend eine hierarchische Ordnung ableiten; die beiden Ziffern nach dem Dezimalpunkt müssen also als Einheit betrachtet werden. So ist z.B. die Klasse „44.43 – Medizinische Mikrobiologie“ keine Unterklasse von „44.40 – Pharmazie, Pharmazeutika“ (Abb. 6).

44.40 Pharmazie, Pharmazeutika

Hier: Apothekenwesen
Verw.: Arzneimittelherstellung -> 58.28 (Pharmazeutische Technologie)
Arzneimittel- und Apothekenrecht -> 86.56 (Gesundheitsrecht, Lebensmittelrecht)

44.41 Pharmazeutische Biologie

Hier: Giftpflanzen; Heilpflanzen

44.42 Pharmazeutische Chemie

44.43 Medizinische Mikrobiologie

Hier: Bakteriologie; Medizinische Virologie
Verw.: Parasitologie -> 44.44 (Parasitologie)

Abb. 6: Ausschnitt medizinischer Klassen der BK [104]

4.2.5 Regensburger Verbundklassifikation

Geschichte und Prinzipien

Die *Regensburger Verbundklassifikation* (RVK) wurde in den 1960er-Jahren entwickelt. Zunächst diente sie als Aufstellungssystematik der Universitätsbibliothek Regensburg. Im Laufe der Jahre übernahmen weitere Bibliotheken die RVK, mittlerweile existieren über 140 Institutionen, welche die RVK anwenden. Sie ist jedoch keine einheitliche Klassifikation, sondern besteht aus insgesamt 33 Fachsystematiken, die einzeln betreut werden.[105]

Aufbau der Notationen

Ähnlich wie in der LCC beginnt in der RVK jede Notation mit zwei lateinischen Buchstaben. Diesen wird eine vierstellige Ziffernfolge angehängt. Anders als die DDC ist die RVK nicht dezimal strukturiert, sodass durch Entfernen der letzten Ziffer einer Notation nicht zwangsläufig die zugehörige Oberklasse ausgedrückt wird.

[104] Ausschnitt aus ebd., S. 66
[105] Vgl. Universitätsbibliothek Regensburg (o.J.)

Medizinische Literatur wird in den Hauptklassen X und Y klassifiziert. Ihrer ursprünglichen Intention als Aufstellungssystematik entsprechend existiert in der RVK eine eigene Systemstelle für medizinische Zeitschriften (XA 10000). Neben dieser existieren 26 weitere Unterklassen auf der zweiten Hierarchieebene (siehe Abb. 7).

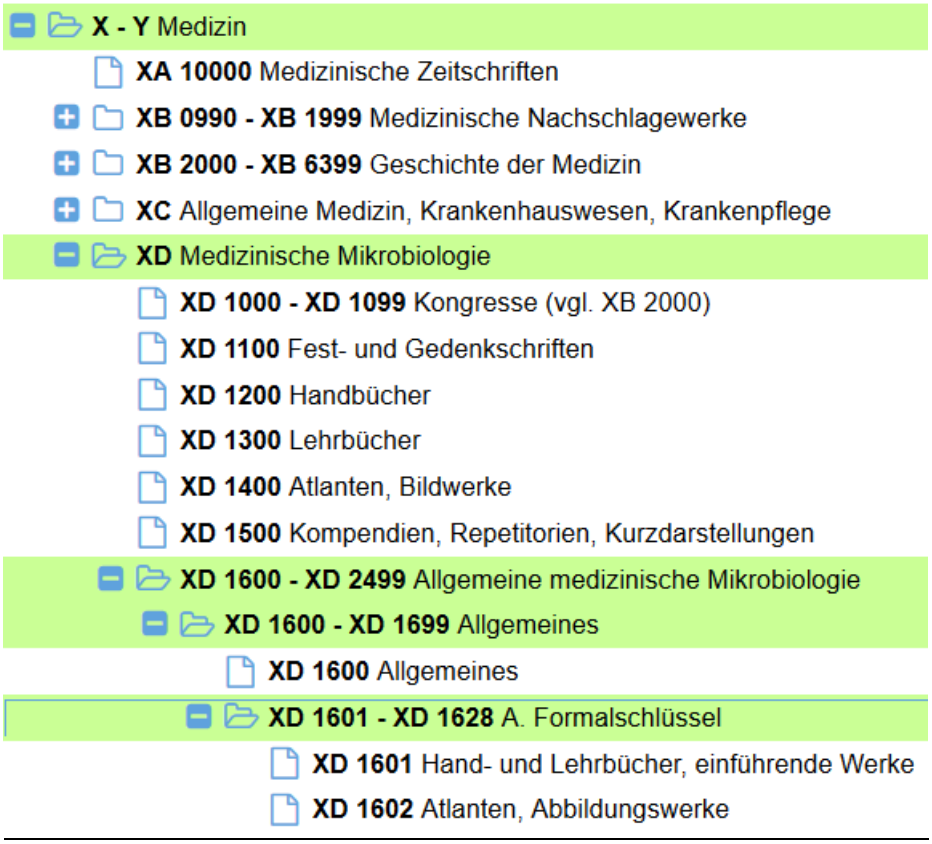

Abb. 7: Ausschnitt medizinischer Klassen der RVK [106]

[106] Screenshot aus der RVK online, verfügbar unter https://rvk.uni-regensburg.de/regensburger-verbundklassifikation-online, zuletzt geprüft am 25.08.2016.

TEIL 2 - UNTERSUCHUNGEN

Entsprechend den in Abschnitt 2.4 vorgestellten Phasen des Data Mining wurden auch die im Rahmen der vorliegenden Arbeit durchgeführten Untersuchungen strukturiert. Sie orientieren sich teilweise an Larson, der seine Forschungen wie folgt beschrieb:

> „The experiments presented in this article return, in some sense, to the early work in automatic classification by attempting to select the correct predefined class based on the characteristics of documents previously assigned to that class. They differ, however, in that the predefined classes used are drawn from the Library of Congress Classification, and the database of previously classified documents used to define the characteristics of the classes consists of over 30,000 MARC records. These experiments differ from most automatic classification studies in that the only content representation used in the classification process is that available in ordinary MARC records, specifically, the titles and subject headings of a given book."[107]

Den Erläuterungen der nachfolgenden Abschnitte seien jedoch bereits zwei wesentliche Unterschiede vorangestellt:

a) In den folgenden Untersuchungen wird nicht die LCC, sondern die NLMC verwendet.
b) Die Dokumente werden durch ihre Zugehörigkeit zu den im ersten Teil genannten Klassifikationen und nicht durch Schlagwörter und Titel repräsentiert.

5 Selektion der Ausgangsdaten

Die Untersuchungen basieren auf Katalogdatensätzen der Bibliothek der Medizinischen Hochschule Hannover. Aus der Datenbank des Gemeinsamen Verbundkatalogs wurden Anfang Juli 2016 sämtliche Einträge heruntergeladen, die zu diesem Zeitpunkt im Feld „Lokale Notation" einen Wert eingetragen hatten. Dadurch lagen alle bis dato intellektuell erschlossenen und mit einer durch die Bibliothek der MHH vergebenen Notation ausgestatteten Datensätze im CSV-Format vor. Insgesamt befanden sich 45.350 Datensätze in dieser CSV-Datei.

5.1 Verteilung der Klassen der NLMC

Eine erste Analyse der Verteilung der Datensätze auf die einzelnen vorhandenen Hauptklassen zeigt, dass diese weit gestreut sind (Tab. 1): Die medizinischen Klassen QS-QZ sowie W-WZ beinhalten zusammen 34.705 (76,53 %) der Datensätze. Insgesamt existieren 4.768 verschiedene Klassen, von denen 2.368 (49,66 %) im medizinischen Bereich und 2.400

[107] Larson (1992), S. 131

(50,34 %) in sonstigen Themenfeldern liegen. Da nur die Klassifizierung der NLMC betrachtet werden soll, wurde der Datenbestand auf die entsprechenden Datensätze reduziert. Die 2.368 übrig gebliebenen unterschiedlichen Klassen sind sehr ungleichmäßig belegt (Abb. 8): 656 Klassen enthalten nur ein Dokument, 323 Klassen enthalten zwei Dokumente und 195 Klassen enthalten drei Dokumente. Somit sind in 1.174 Klassen (49,58 %) maximal drei Datensätze enthalten. Die 24 meistbelegten Klassen enthalten zusammen 7.774 aller Dokumente, sodass das Prozent der größten Klassen 22,37 % aller Dokumente enthält. Man spricht in diesem Fall von einer verzerrten Verteilung („skewed distribution“)[108].

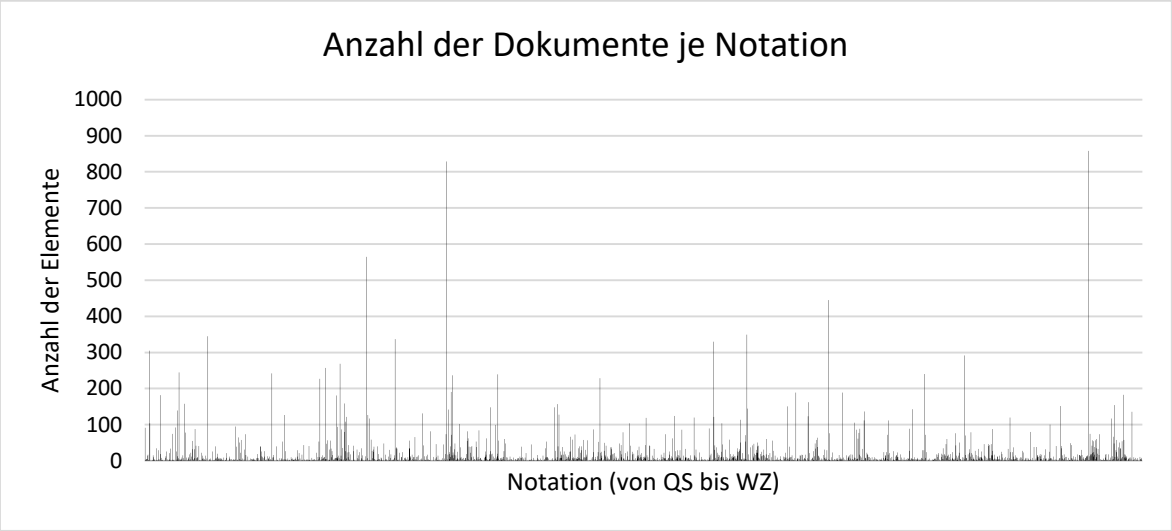

Abb. 8: Anzahl der Dokumente je vergebener Notation

Die Betrachtung der auf die Anzahl der Dokumente bezogenen größten und der kleinsten Klassen spiegelt aufschlussreiche Informationen über den Datenbestand wider (Tab. 2 und Tab. 3). Die größten Klassen sind häufig die übergeordneten Hauptklassen; nur wenige „detaillierte“ Klassen werden besonders häufig vergeben. Auch diese decken allerdings i.d.R. allgemeinere Themen ab.[109] Dementsprechend sind die meisten der betrachteten Dokumente Serien bzw. Zeitschriften (welche von der Bibliothek der MHH ja in den Hauptklassen erschlossen werden), bzw. allgemeine Literatur wie z.B. Lehrbücher. Die Daten in Tab. 3 verdeutlichen beispielhaft die Verwendung der „Table G“ in der NLMC: Durch Hinzufügen einer Buchstaben-Zahlen-Kombination an eine Notation wird der Inhalt einer Publikation auf einen geographischen Bereich beschränkt. Hierdurch ergeben sich natürlich in manchen Klassen viele Varianten, die aber letztendlich auf einer gemeinsamen Notation basieren. Ein automatisches System würde dies jedoch nicht erkennen und jede dieser Varianten als eigene

[108] Wang (2009), S. 2269
[109] W 50 beinhaltet bspw. Literatur zur medizinischen Ethik, QS 4 allgemeine anatomische Literatur (z.B. Lehrbücher) und QZ 200 allgemeine Werke über Tumore.

Kategorie ansehen. Daher ist es in der Phase der Datenvorverarbeitung nötig, solche Klassen zusammenzuführen, um eine sinnvolle Auswertung mit Data-Mining-Verfahren möglich zu machen.

Tab. 1: Verteilung der Datensätze auf die Hauptklassen der NLMC

A	414
B	1535
C	257
D	749
E	5
F	7
G	407
H	1276
J	217
K	250
L	604
M	9
N	38
P	545
Q-QR	2646
QS-QZ	6206
S	94
T	248
U	36
V	2
W	28499
X	1306
Gesamt	**45350**

Tab. 2: Die 20 am häufigsten vergebenen Notationen

WZ	858
WB	829
W 50	565
WO	445
WM	349
QV	345
WA	337
WL	330
QS 4	305
WU	292
W	269
QZ 200	257
QU	245
QW	242
WS	240
WB 930	239
WB 115	237
WG	229
QZ	227
WB 105	191

Tab. 3: Klassen mit nur einem Dokument (Auswahl)

WZ 308	1
WZ 32	1
WZ 39	1
WZ 70 DA1	1
WZ 70 DA15	1
WZ 70 GA785	1
WZ 70 GC8	1
WZ 70 GG 4	1
WZ 70 GG6	1
WZ 70 GH8	1
WZ 70 GN4	1
WZ 70 GS9	1
WZ 70 HA1	1
WZ 70 HN5	1
WZ 70 HT3	1
WZ 70 JJ3	1
WZ 70 JT5	1
WZ 70 JT8	1
WZ 80	1
WZ 9 D	1

Die bereits angesprochene Tatsache, dass knapp die Hälfte aller vergebenen Klassen maximal drei Dokumente beinhalten, wird als *data sparseness* bezeichnet: „Bibliographic corpora are very sparse. There are a large number of categories with an insufficient number of documents.“[110] Dies ist eine Charakteristik vieler in der Praxis vorkommender Datenbestände. Um einen Klassifikator jedoch effizient trainieren zu können, muss eine bestimmte Zahl von Dokumenten pro Klasse gegeben sein:

> „It is clear that the average document quantity per category follows the power-law distribution. The power law says that there are a large number of categories with a few documents assigned to them, but a small number of categories holding a large percent-

[110] Wang (2009), S. 2271

> age of the documents in the corpus. This imposes great difficulty upon supervised learning algorithms, because the inductive construction of text classifiers usually assumes sufficient training instances and even distribution of categories."[111]

Je mehr Kategorien vorhanden sind und je weniger Dokumente einer Klasse angehören, desto schwieriger ist es, anhand dieser wenigen Beispieldokumente die Charakteristika einzelner Klassen zu lernen. Gerade minimale Unterschiede zwischen Klassen können durch data sparseness kaum ermittelt werden.[112]

5.2 Attributauswahl

Die aus dem Gesamtkatalog heruntergeladenen Datensätze wurden in einem zweiten Schritt gemäß der Fragestellung der vorliegenden Arbeit aufbereitet. Sämtliche Attribute bis auf die Zugehörigkeiten zu den im ersten Teil vorgestellten Klassifikationen wurden entfernt. Ein Datensatz besteht jetzt noch aus folgenden Attributen: der PPN (einem individuellen Identifikator), der/den zugeordneten Klasse(n) der DDC, NLM, RVK, LCC, BK und der von der Bibliothek der MHH vergebenen Notation der NLMC („Lokale Notation"). Beispiele sind in Tab. 4 dargestellt. Bei fast allen Datensätzen bestehen Mehrfachbelegungen in einem oder mehreren Attributen. Der Datensatz mit der PPN 477072496 besitzt z.B. drei unterschiedliche DDC-Notationen, zwei NLMC-Notationen, eine LCC-Notation und zwei BK-Notationen, jeweils durch Semikolon getrennt. Diese Eigenschaft ist bei der folgenden Datenvorverarbeitung zu berücksichtigen.

Neben der in Abschnitt 5.1 dargestellten verzerrten und verstreuten Verteilung der Datensätze sowie der Mehrfachbelegung mancher Felder ist insbesondere zu beachten, dass einige Datensätze fehlerhafte Eintragungen wie z.B. unnötige Leerzeichen, falsche Feldbelegungen (z.B. wenn die Notation der BK im Feld für die RVK eingetragen ist) oder Rechtschreibfehler enthalten. Die in Tab. 3 vorhandene Klasse „WZ 70 GG 4" enthält nur ein Dokument; allerdings existiert eine weitere, korrekt vergebene Notation „WZ 70 GG4", die 136 Dokumente enthält. Im späteren Lernprozess würden diese beiden eigentlich identischen Klassen aufgrund des überflüssigen Leerzeichens als unterschiedliche Kategorien interpretiert werden. Daher sind solche Fehler vor der Anwendung von Data-Mining-Verfahren möglichst zu beheben.

[111] Ebd., S. 2272
[112] Vgl. ebd.

Tab. 4: Beispieldatensätze mit ausgewählten Attributen

PPN	DDC	NLM	RVK	LCC	BK	Lokale Notation
556727703	610				44.34$jAnatomie	QS
524230250	590; 610				44.34$jAnatomie	QS
501082719	610; 610				44.87$jGastroenterologie	QS
497326523	571.64; 611.0181				42.15$jZellbiologie; 44.35$jHistologie$XMedizin	QS
49653419X	616.0472; 610				44.90$jNeurologie	QS
496533711	612.022; 570		WD 8050	QP383.8	44.90$jNeurologie	QS
49433570X	573.869			QL801; QM465	44.90$jNeurologie; 42.15$jZellbiologie	QS
48361923X	571; 571.845			QL801	44.88$jUrologie$jNephrologie	QS
482764546	616.856				44.90$jNeurologie	QS
477072496	571; 612.75; 599.9	W1; QS 532.5.C7		QL801	42.15$jZellbiologie; 44.83$jRheumatologie$jOrthopädie	QS
379615932	571			QL801; QM431	44.36$jEmbryologie$XMedizin; 44.83$jRheumatologie$jOrthopädie	QS
374169721	571; 617.954		WA 11700; WX 6600	QL801; QH499	42.15$jZellbiologie	QS

Nicht jeder Datensatz ist darüber hinaus mit jeder Klassifikation erschlossen. Abb. 9 zeigt auf, wie viele der 34.705 untersuchten Datensätze einen Wert im Attributfeld der jeweiligen Klassifikation besitzen.[113]

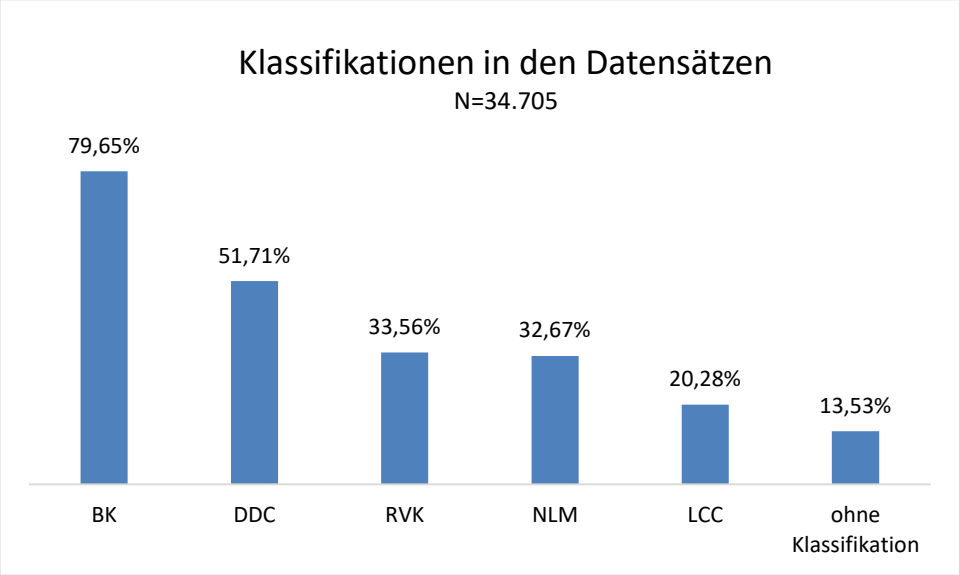

Abb. 9: Relative Häufigkeit der mit den Klassifikationen erschlossenen Datensätze

Die 4.695 Datensätze (13,53 %), die lediglich eine PPN und Lokale Notation – aber keine weitere Klassifikation – aufweisen, sind nicht für die Erstellung eines Klassifikators geeignet, da sie keinerlei zu lernende Informationen besitzen. Daher werden diese Datensätze aus dem Gesamtdatenbestand entfernt, sodass schließlich noch eine Ausgangsdatenbasis von 30.010 Datensätzen vorhanden ist. Diese ist das Fundament der folgenden Data-Mining-Untersuchungen. Die wesentlichen Eigenschaften der bisher betrachteten Katalogdaten sind in Tab. 5 noch einmal zusammenfassend dargestellt.

Tab. 5: Zusammenfassende Übersicht der untersuchten Datensätze

Von der MHH-Bibliothek erschlossene Datensätze	45.350
… davon im Bereich QS-QZ oder W-WZ	34.705
… … davon mit zusätzlicher Klassifikation	30.010
… … … davon mit DDC	17.946
… … … davon mit NLM	11.338
… … … davon mit RVK	11.646
… … … davon mit LCC	7.039
… … … davon mit BK	27.643

[113] Das Feld „Lokale Notation“ ist selbstverständlich bei allen Datensätzen belegt.

6 Datenvorverarbeitung

Um die im vorherigen Abschnitt angesprochenen Probleme des Datenbestandes möglichst vollständig zu beheben, wurden anschließend an die ersten Analysen die Attribute bereinigt. Dieser Prozess wird für jede Klassifikation einzeln dargestellt. Die notwendigen Schritte richten sich nach der in Abschnitt 4.2 vorgestellten Struktur ordentlicher Notationen. Im Anschluss an die einzelnen Schritte wird jeweils ein Beispiel gegeben, wie das Attribut vor und nach der Bereinigung ausgeprägt war.

6.1 LCC

a) Bei Datensätzen mit mehr als einer LCC-Notation wurde nur die erste berücksichtigt, alle weiteren vergebenen Notationen wurden entfernt (BF575.A6; RC531 → BF575.A6). Oftmals war dieselbe Notation doppelt vergeben, sodass hierdurch kein Verlust entstand (z.B. KKW62.3; KKW62.3 → KKW62.3).

b) Bei Datensätzen mit Kleinbuchstaben im Attributwert wurde dieses Feld leer gelassen (med 001b → <leer>). Es handelt sich hierbei vermutlich um falsche Feldbelegungen. Dabei gehen auch potentiell brauchbare Informationen verloren (QL801.E67 vol. 160 → <leer>), allerdings wäre es zu aufwendig, all diese Möglichkeiten manuell zu kontrollieren.

c) Datenfelder mit mehr als drei aufeinanderfolgenden Großbuchstaben wurden ebenfalls entfernt (IN PROCESS (ONLINE) → <leer>), da die LCC für reguläre Notationen maximal drei Großbuchstaben nacheinander zulässt.

d) Sonderzeichen wurden entfernt (RM214.5+ → RM214.5).

e) Bei der Angabe von Notationsbereichen wurde die erste Notation des Bereiches als Attributwert angenommen (QR180-189.5 → QR180).

f) Leerzeichen, die vor oder nach einem Dezimalpunkt stehen, wurden entfernt (RC86.7 .D36 2010 → RC86.7.D36 2010 bzw. RC86.7. F74 2011 → RC86.7.F74 2011).

g) Zum Schluss wurden Jahreszahlen am Ende einer Notation sowie Buchstaben-Zahlen-Kombinationen nach einem Dezimalpunkt entfernt, da es sich hierbei um Elemente der book number oder alphabetische Ordnungskriterien handelt (KK8494.E38 2001 → KK8494), siehe auch Abschnitt 4.2.1 dieser Arbeit.

Vor der Datenaufbereitung existierten 3.050 unterschiedliche Werte (von denen nicht alle eine wirkliche Klasse der LCC darstellten, siehe z.B. Punkt c)), nach den o.g. Schritten nur noch 2.037 unterschiedliche LCC-Klassen.

6.2 DDC

a) Analog zum Vorgehen bei der LCC wurden auch bei der DDC nur die erstgenannten Notationen berücksichtigt (930; 610 → 930).
b) Da die Notationen der DDC nur aus Zahlen und evtl. Dezimalpunkten bestehen, wurden alle Werte entfernt, die einen Buchstaben enthielten (884 (HIPPO) A40 → <leer>).
c) Datenfelder mit Sonderzeichen oder nicht in der DDC vorgesehenen Zeichen wurden soweit möglich bereinigt bzw. entfernt (617.95405$2DDC22ger → 617.095405 bzw. 617'.95'00924 → 617.9500924).
d) Notationen mit weniger als drei Ziffern wurden entfernt (33 → <leer>).
e) Notationen mit mehr als drei Ziffern und einer „0“ als letzter Ziffer wurden um diese „0“ gekürzt.
f) Dezimalpunkte wurden entfernt, da sie nur der Übersichtlichkeit dienen (616.96 → 61696).

Von vorher 5.723 unterschiedlichen Werten sind nach der Bereinigung nur noch 3.035 vorhanden.

6.3 NLMC

a) Bei der NLMC wurden ebenfalls nur die erstgenannten Notationen verwendet (QH 585.2; QH 581.2 → QH 585.2).
b) Bei Datensätzen, deren Attributwert mit etwas anderem als einem Großbuchstaben anfängt (z.B. Jahreszahlen), wurde dieser Wert entfernt, da dies nicht dem Aufbau der NLMC-Notationen entspricht (1994 F-920 → <leer>).
c) Cutter-Nummern, Kombinationen aus Cutter-Nummer und Veröffentlichungsjahr sowie Angaben aus der „Table G“ wurden entfernt (QW 11.1 M643 → QW 11.1 bzw. QS 17 R737c 1998 → QS 17 bzw. QW 575.5.A6 → QW 575.5).
d) Fehlende Leerzeichen zwischen Buchstaben und Ziffern wurden ergänzt (WZ309 → WZ 309).
e) Sonstige fehlerhafte Einträge (z.B. fehlende oder überflüssige Leerzeichen an anderer Stelle der Notation) wurden nach Möglichkeit bereinigt, ansonsten entfernt (QS 532.5M5344 → QS 532.5 oder W1UE v.17 2011 → W1).

Von vorher 2.662 unterschiedlichen Werten sind nach der Bereinigung nur noch 1.612 vorhanden.

6.4 BK

Für die BK waren vor der Datenbereinigung 5.535 unterschiedliche Einträge vorhanden. Dieser hohe Wert liegt vor allem darin begründet, dass i.d.R. mehrere Klassen der BK vergeben werden und jede dieser Kombinationen als individueller Wert interpretiert wird. So finden sich z.B. die folgenden beiden Einträge, die sich nur im Grad ihrer Detailliertheit unterscheiden:

02.01$jGeschichte der Wissenschaft und Kultur; 44.01$jGeschichte der Medizin

02.01$jGeschichte der Wissenschaft und Kultur; 44.01$jGeschichte der Medizin; 44.65$jChirurgie

Datensätze, die mit einer dieser BK-Kombinationen erschlossen sind, sind sich recht ähnlich, würden von einem automatischen Verfahren aufgrund ihrer Nicht-Identität jedoch als komplett unterschiedlich interpretiert werden. Die Ähnlichkeit ist nur für einen menschlichen Betrachter erkennbar. Um eine für den Computer erkenn- und berechenbare Ähnlichkeit zwischen den Datensätzen herzustellen, wurden für die vorliegende Untersuchung sämtliche Daten zur BK in Vektoren transformiert. Dies wird in Abschnitt 7 näher erläutert. Zur Vorbereitung wurden an dieser Stelle zunächst alle Informationen bis auf die Notationen aus den Einträgen zur BK entfernt (44.01$jGeschichte der Medizin → 44.01). Darüber hinaus wurden die wenigen vorhandenen fehlerhaften Einträge bereinigt bzw. entfernt.

6.5 RVK

a) Wie schon bei den o.g. Klassifikationen wurde auch bei der RVK jeweils nur die erstgenannte Notation berücksichtigt (ZY 4850; ZY 4852 → ZY 4850).
b) Einträge, die mit „ELIB“ beginnen, wurden entfernt, da diese nicht den Notationen der RVK entsprechen und daher offenbar andere Sachverhalte abbilden.
c) Fehlende Leerzeichen wurden ergänzt (YR1905 → YR 1905).
d) Weitere fehlerhafte Einträge wurden entfernt (z.B. G:at S:hh → <leer>).

Nach der Bereinigung sind von vorher 5.816 unterschiedlichen Einträgen im Feld RVK nur noch 3.562 vorhanden.

6.6 Lokale Notation

Bei der Bereinigung der von der MHH vergebenen Notationen wurde analog zu Abschnitt 6.3 vorgegangen. Von zunächst 2.368 unterschiedlichen Einträgen sind nach der Bereinigung noch 1.935 vorhanden. Werden darüber hinaus Unterteilungen mittels Dezimalpunkten

ignoriert, so ergeben sich nur 1.730 unterschiedliche Werte. Um im späteren Verlauf diese beiden Grade der Bereinigung miteinander vergleichen zu können, wird den Datensätzen ein weiteres Attribut hinzugefügt, welches die stärker „gekürzten“ Notationen enthält.

6.7 Auswertung und Aufbereitung

Nach der Datenvorverarbeitung enthält ein Datensatz nun also folgende Attribute: PPN, DDC, NLM, RVK, LCC, BK, die lokale Notation und die gekürzte lokale Notation. Dies kann z.B. so aussehen:

524950679 | 6168917 | WM 420 | CU 8400 | RC475 | 44.91; 77.77 | WM 420.5 | WM 420

Es bleiben insgesamt 29.946 Datensätze übrig, die neben der lokalen Notation mit mindestens einer weiteren Klassifikation erschlossen sind. Tab. 6 vergleicht den bereinigten Datenbestand mit den Angaben aus Tab. 5.

Tab. 6: Übersicht über die bereinigten Datensätze

	vor Bereinigung (s. auch Tab. 5)	nach Bereinigung
Von der MHH-Bibliothek erschlossene Datensätze	45.350	45.350
… davon im Bereich QS-QZ oder W-WZ	34.705	34.705
… … davon mit zusätzlicher Klassifikation	30.010	29.946 (100 %)
… … … davon mit DDC	17.946	17.501 (58,44 %)
… … … davon mit NLM	11.338	10.890 (36,37 %)
… … … davon mit RVK	11.646	11.377 (37,99 %)
… … … davon mit LCC	7.039	6.926 (23,13 %)
… … … davon mit BK	27.643	27.643 (92,31 %)

Durch die Datenvorverarbeitung hat sich auch der Anteil der nur mit wenigen Dokumenten belegten Klassen verändert. Tab. 7 stellt die Ausprägungen der *data sparseness* vor und nach der Datenbereinigung gegenüber. Es zeigt sich, dass der Anteil der Dokumente, die in den größten Klassen eingeordnet wurden, in etwa gleich geblieben ist. Hingegen konnte der Anteil an schwach belegten Klassen von 49,58 % auf 45,63 % bzw. 44,16 % verringert werden. In den schwach belegten Klassen sind vorher 6,23 % und hinterher 4,94 % bzw. 4,32 % der Dokumente enthalten. Allerdings sind auch dies immer noch Werte, bei denen die sinnvolle Anwendung (bestimmter) automatischer Lernverfahren Schwierigkeiten bereiten kann.

Tab. 7: Data sparseness vor und nach der Datenvorverarbeitung

	vorher	nachher	nachher (bei gekürzter Notation)
Anzahl der Klassen	2.368	1.935	1.730
Klassen mit…			
… 1 Dokument	656	458	388
… 2 Dokumenten	323	254	223
… 3 Dokumenten	195	171	153
… max. 3 Dokumenten	1.174 (49,58 %)	883 (45,63 %)	764 (44,16 %)
Dokumente in Klassen mit max. 3 Dokumenten	1.887 (6,23 %)	1.479 (4,94 %)	1.293 (4,32 %)
Dokumente in den 24 größten Klassen	7.774 (22,37 %)	7.049 (23,54 %)	7.149 (23,87 %)

Tab. 8: Die zehn größten Klassen (nach Datenvorverarbeitung)

Klasse	**Anzahl der Dokumente**
WZ	783 (2,61 %)
WB	686 (2,29 %)
W 50	545 (1,82 %)
WZ 100	443 (1,48 %)
WO	355 (1,19 %)
QV	322 (1,08 %)
WA	305 (1,02 %)
WL	304 (1,02 %)
WM	303 (1,01 %)
WU	258 (0,86 %)

Die zehn nach der Datenvorverarbeitung am häufigsten vergebenen Klassen sind in Tab. 8 dargestellt. Hier zeigt sich, dass weiterhin die Hauptklassen, in denen vorrangig Zeitschriften und Serien erschlossen werden, am stärksten belegt sind. Um aussagekräftige und in der Praxis verwendbare Ergebnisse zu erhalten, wird aufgrund dieser Tatsache der Datenbestand noch ein weiteres Mal deutlich verringert.

a) Zunächst werden sämtliche Datensätze entfernt, die mit einer Hauptklasse erschlossen sind. Dies betrifft 6.475 (21,62 %) der 29.946 Datensätze. Hierdurch wird also

ein nicht unerheblicher Teil der Datensätze aus dem Data-Mining-Verfahren herausgenommen. Dies ist allerdings vertretbar, da es zu keinem Widerspruch mit den Anforderungen der praktischen Sacherschließung kommt: Zeitschriften und Serien sind nämlich erstens seltener neu zu erschließen und können zweitens i.d.R. deutlich einfacher und schneller einem Themenbereich zugeordnet werden. Bei diesen Ressourcen ist eine automatische Unterstützung durch den Computer also nicht unbedingt vonnöten.

b) Darüber hinaus werden diejenigen Datensätze entfernt, die in Klassen erschlossen sind, in denen sich insgesamt weniger als zehn Dokumente befinden. Dies betrifft 4.143 Datensätze. Bei dem Großteil dieser Datensätze ist erstens davon auszugehen, dass sie auch in anderen, häufiger vergebenen Klassen treffend verortet werden können; zweitens werden Klassen, die bisher in der MHH nur wenige Male vergeben wurden, in Zukunft wahrscheinlich auch selten (oder sogar nie) verwendet werden. Insgesamt verbleiben jetzt noch 19.328 Datensätze im Datenbestand.

Die wesentlichen Eigenschaften des umfassend aufbereiteten Datenbestandes finden sich in Tab. 9 bis Tab. 12. Aufgrund der Tatsache, dass die gekürzten Notationen zwangsläufig weniger Klassen zur Folge haben, ist die durchschnittliche Zahl der Dokumente pro Klasse bei diesem Datenbestand größer. Beide Datenbestände sind weiterhin recht ungleich verteilt; mit wenigen Klassen, die einen Großteil der Dokumente beinhalten und vielen Klassen, denen nur wenige Dokumente zugeordnet sind. Tab. 12 veranschaulicht den jeweiligen Anteil an den Gesamtdaten, der mit nur einer Klassifikation erschlossen ist.

Tab. 9: Übersicht über den aufbereiteten Datenbestand

	LokNot_ganz	LokNot_kurz
Anzahl der Dokumente	19.328	19.328
Anzahl der Klassen	568	514
Durchschn. Anzahl an Dokumenten pro Klasse	34 Dokumente	37,6 Dokumente
Dokumente in den 20 größten Klassen	3.770 (19,51 %)	3.972 (20,55 %)

Tab. 10: Die zehn größten Klassen (LokNot_ganz)

Klasse	Anzahl der Dokumente
W 50	545 (2,82 %)
WZ 100	443 (2,29 %)
QS 4	233 (1,21 %)
QZ 200	223 (1,15 %)
WZ 70	202 (1,05 %)
WB 115	174 (0,90 %)
WB 105	160 (0,83 %)
W 18.2	158 (0,82 %)
WO 200	154 (0,80 %)
WB 930	153 (0,79 %)

Tab. 11: Die zehn größten Klassen (LokNot_kurz)

Klasse	Anzahl der Dokumente
W 50	545 (2,82 %)
WZ 100	443 (2,29 %)
QS 4	233 (1,21 %)
W 18	229 (1,18 %)
QZ 200	223 (1,15 %)
WZ 70	202 (1,05 %)
WB 115	174 (0,90 %)
W 20	164 (0,85 %)
WM 420	163 (0,84 %)
WB 105	160 (0,83 %)

Tab. 12: Datensätze mit nur einer Klassifikation

Klassifikation	Datensätze
nur DDC	669 (3,46 %)
nur LCC	37 (0,19 %)
nur NLMC	333 (1,72 %)
nur RVK	53 (0,27 %)
nur BK	3.835 (19,84 %)

7 Transformation

Für die Durchführung der Data-Mining-Verfahren wird das lizenzfreie und quelloffene Programm WEKA verwendet, das von der University of Waikato entwickelt wurde.[114] Dieses benötigt die Eingabedaten im ARFF-Format[115], das speziell für WEKA zugeschnitten ist. Daher wird die bislang als CSV-Datei vorliegende Datenbasis nun in das ARFF-Format transformiert.

Eine ARFF-Datei ist in zwei Teile gegliedert: Der Kopf enthält den Namen des Datenbestandes (eingeleitet mit *@relation*) und die Angaben über Name und Typ der verwendeten Attribute (für jedes Attribut eingeleitet durch *@attribute*). Der Hauptteil beginnt mit einer

[114] Download und weitere Informationen unter http://www.cs.waikato.ac.nz/ml/weka/, zuletzt geprüft am 10.09.2016.
[115] ARFF steht für *Attribute-Relation File Format*

Zeile, auf der *@data* steht, worauf in den folgenden Zeilen jeweils ein Datensatz folgt. Als Trennzeichen zwischen den Attributen eines Datensatzes verwendet ARFF Kommata, leere Werte müssen mit einem *?* gekennzeichnet werden. Abb. 10 zeigt den Beginn der zunächst erstellten ARFF-Datei. Zur einfacheren Verarbeitung und da es für den maschinellen Lernprozess keinen Unterschied macht, wurden sämtliche noch vorhandene Leerzeichen innerhalb der Notationen durch Unterstriche ersetzt (siehe z.B. Zeile 13 in Abb. 10). Als Datentyp wurde für alle Attribute zunächst *string* angenommen, da es sich bei allen verwendeten Attributen um Zeichenketten handelt.

```
1  @relation datamining-mhh
2
3  @attribute PPN string
4  @attribute DDC string
5  @attribute NLM string
6  @attribute RVK string
7  @attribute LCC string
8  @attribute BK string
9  @attribute LokNot_ganz string
10 @attribute LokNot_kurz string
11
12 @data
13 739225030,610,WB_18.2,WW_1452,?,44.34,QS_18.2,QS_18
14 631566821,610,QS_18.2,WW_1454,?,44.34,QS_18.2,QS_18
15 556736028,610,QS_4,WW_1454,?,44.34,QS_18.2,QS_18
16 359632289,?,WB_18.2,WW_1456,?,44.34,QS_18.2,QS_18
17 66173384X,610,QS_4,XF_1216,?,44.34,QS_18.2,QS_18
18 668385499,?,WB_18.2,XF_1216,?,44.34,QS_18.2,QS_18
19 604281048,?,WB_18.2,XF_1216,?,44.34,QS_18.2,QS_18
20 560542755,?,WB_18.2,XF_1216,?,44.34,QS_18.2,QS_18
```

Abb. 10: Datensätze im ARFF-Format

Nachdem diese Datei in WEKA eingelesen wurde, können mit Hilfe dort vorhandener Werkzeuge („Filter“) weitere Anpassungen vorgenommen werden. Diese sind nötig, um die Ausgangsdaten für die Anwendung von Data-Mining-Verfahren weiter aufzubereiten.

a) Zuerst wird das Attribut PPN entfernt, da dieses später nicht in den automatischen Lernprozess mit einbezogen werden soll. In den Ausgangsdaten wurde es bisher beibehalten, um die einzelnen Datensätze eindeutig identifizieren zu können. (Dies ist u.a. nötig, um in einem erneuten Durchlauf der bisherigen Prozessschritte geänderte oder zusätzliche Daten in die Datenbasis integrieren zu können.)
b) Alle anderen Attribute bis auf die Basisklassifikation wurden vom Datentyp *string* zu *nominal* konvertiert. Hierdurch betrachtet WEKA den konkreten Wert eines dieser Attribute als Element aus der Menge aller für dieses Attribut in den Datensätzen

vorhandenen Werte. Diese Menge wird durch WEKA in der Attributbeschreibung vollständig angefügt (siehe Abb. 11).

c) Das Attribut BK wurde vom Datentyp *string* zu einem sog. *word vector* konvertiert. Hierfür betrachtet WEKA alle vorkommenden Werte für das Attribut BK und fügt diese jeweils als eigenständiges Attribut dem Datenbestand hinzu. Für jedes dieser neuen Attribute wird in den Datensätzen nun abgebildet, ob es dort enthalten ist oder nicht. Ein Datensatz, der z.B. mit den BK-Klassen 44.01 und 44.12 erschlossen ist, erhält bei diesen beiden Attributen den Wert 1 und für alle anderen aus der BK generierten Attribute den Wert 0.

Abb. 11 verdeutlicht die Auswirkungen der in diesem Schritt durchgeführten Anpassungen. Insgesamt enthält die Datei jetzt 662 Attribute. Vier Attribute repräsentieren Klassifikationen, ein Attribut bildet die von der MHH-Bibliothek vergebene Notation ab und die restlichen 657 Attribute stehen jeweils für eine Klasse der BK.

```
1  @relation 'datamining-mhh-weka.filters.unsupervised.attribute.Remove-R
2
3  @attribute DDC {610,5108,6120076,611,61100222,61100223,6110022,6110222
4  @attribute NLM {WB_18.2,QS_18.2,QS_4,WZ_290,QS_17,QS_25,WE_101,WY_100,
5  @attribute RVK {WW_1452,WW_1454,WW_1456,XF_1216,ZX_9860,WW_1000,WW_145
6  @attribute LCC {RD32,QP29,QM1,QM23.2,QM24,QM25,QM30,QM531,RT1,QM21,QP3
7  @attribute LokNot_ganz {QS_18.2,QS_4,QS_504,QS_525,QS_604,QS_675,QT_10
8  @attribute 44.34 numeric
9  @attribute 44.65 numeric
10 @attribute 08.24 numeric
11 @attribute 35.70 numeric
12 @attribute 42.88 numeric
13 @attribute 44.01 numeric
14 @attribute 44.06 numeric
```

```
666 @data
667 {3 ?,5 1}
668 {1 QS_18.2,2 WW_1454,3 ?,5 1}
669 {1 QS_4,2 WW_1454,3 ?,5 1}
670 {0 ?,2 WW_1456,3 ?,5 1}
671 {1 QS_4,2 XF_1216,3 ?,5 1}
672 {0 ?,2 XF_1216,3 ?,5 1}
673 {0 ?,2 XF_1216,3 ?,5 1}
674 {0 ?,2 XF_1216,3 ?,5 1}
675 {0 ?,2 XF_1216,3 ?,5 1}
676 {0 ?,2 XF_1216,3 ?,5 1}
677 {0 ?,2 ?,3 ?,5 1}
678 {1 ?,2 ?,3 ?,5 1}
```

Abb. 11: Datensätze nach Anwendung von WEKA-Filtern

8 Data Mining

Nach den umfangreichen Vorbereitungen wird im folgenden Schritt das eigentliche maschinelle Lernen angewendet. Da das Ziel ist, durch die Analyse bereits erschlossener Datensätze die Attributwerte neuer Datensätze zu bestimmen, handelt es sich hierbei um überwachtes Lernen, genauer gesagt um Klassifizierung (siehe auch Abschnitt 2.3.2 dieser Arbeit). Die Data-Mining-Verfahren werden zweimal angewendet: beim ersten Durchlauf mit der ungekürzten Notation (*LokNot_ganz*) als zu lernende Klasse, beim zweiten Mal mit der gekürzten Notation (*LokNot_kurz*) als Zielklasse.

Die Unterteilung der Datenmenge in Trainings- und Testdokumente wird nicht manuell vorgenommen. Stattdessen wird der gesamte Datenbestand genutzt und zur Evaluierung die in WEKA vorhandene Möglichkeit der *cross-validation* verwendet. Hierbei wird der Lernprozess mehrfach durchlaufen und jedes Mal die Menge der Trainings- bzw. Testdokumente verändert. Dadurch ergeben sich (je nach Einstellung) z.B. 10 Evaluierungsergebnisse, die anschließend zusammengeführt und gemeinsam ausgewertet werden. Dies erhöht durch Einbezug sämtlicher Dokumente in den Lernprozess die Genauigkeit des Verfahrens.[116]

8.1 k-nearest-neighbours (KNN)

Zunächst wird das in dieser Arbeit verwendete Verfahren näher vorgestellt: KNN zählt zu den Verfahren des *Instance-based learning*. Diese Data-Mining-Methoden zeichnen sich dadurch aus, dass in der Trainingsphase lediglich sämtliche Trainingsdokumente gespeichert werden. In der Testphase werden die zu klassifizierenden Dokumente dann auf Ähnlichkeit mit den Testdokumenten geprüft. Grundlage hierfür ist ein vorher definiertes Abstandsmaß. Aufgrund der Tatsache, dass die eigentliche Rechenarbeit in die Testphase verlegt wird, spricht man auch von *lazy classifiers*.[117]

Das ähnlichste Dokument wird auch als „nächster Nachbar“ bezeichnet. Es ist möglich, für die Berechnung der zu findenden Klassen mehrere nächste Nachbarn einzubeziehen. Hierdurch kann der Einfluss eventueller Ausreißer in den Daten eingegrenzt werden – dennoch

[116] Würde hingegen vor Anwendung des Data-Mining-Verfahrens manuell eine Teilmenge von bspw. einem Drittel zur späteren Evaluierung entnommen, so hätte das Verfahren gar nicht die Möglichkeit, seltene Klassen, die dann zufälligerweise nur in den Testdokumenten enthalten sind, in der Trainingsphase zu lernen. Vgl. hierzu auch Witten u.a. (2011), S. 152-154.
[117] Vgl. Witten u.a. (2011), S. 131f. und S. 472

bedeutet die Berücksichtigung mehrerer nächster Nachbarn nicht automatisch eine Verbesserung der Klassifizierungsqualität. Die Anzahl der nächsten Nachbarn wird mit der Variable *k* bezeichnet.

Zur Bestimmung des Abstandes zweier Dokumente existieren verschiedene Maße. Bei der *euklidischen Distanz* wird die Quadratwurzel aus der Summe aller Quadrate der Attributwertdifferenzen berechnet:

$$\sqrt{\left(a_1^{(1)} - a_1^{(2)}\right)^2 + \left(a_2^{(1)} - a_2^{(2)}\right)^2 + \ldots + (a_y^{(1)} - a_y^{(2)})^2}$$

Hierbei besitzt das erste Dokument die Attributwerte $a_1^{(1)}, a_2^{(1)}, \ldots, a_y^{(1)}$ (mit *y* = Anzahl der Attribute) und das zweite Dokument die Werte $a_1^{(2)}, a_2^{(2)}, \ldots, a_y^{(2)}$.[118]

In der vorliegenden Untersuchung wurden folgende WEKA-Einstellungen verwendet:

- IBk-Classifier
- Anzahl der nächsten Nachbarn k = 1
- Abstandsmaß: euklidische Distanz
- Zielklasse: LokNot_ganz bzw. LokNot_kurz
- Cross validation mit 10 folds

8.2 Ergebnisse

Die wesentlichen Ergebnisse sind in Tab. 13 dargestellt: Bei Berücksichtigung der ungekürzten Notation konnten 58,07 % der Testdokumente in die korrekte Klasse eingeordnet werden. Bei Verwendung der um alle Angaben nach einem eventuellen Dezimalpunkt gekürzten Notation ergeben sich 59,40 % korrekte Zuordnungen. Hierbei handelt es sich um diejenigen Datensätze, deren ermittelte Notation exakt der von der MHH-Bibliothek vergebenen Notation entspricht.

Die Klassifizierungsfehler lassen sich in zwei grundlegende Klassen einteilen:

a) Entweder handelt es sich bei der ermittelten Notation um eine alternative oder ähnliche Klasse zur korrekten Klasse, oder
b) die berechnete Klasse hat keine Ähnlichkeit mit der korrekten Klasse.

[118] Vgl. ebd., S. 131

Da eine detaillierte Analyse dieser beiden Fehler voraussetzt, dass man die Ähnlichkeit oder Alternative zweier Klassen vorher zufriedenstellend bestimmt, wurden im Rahmen der hier durchgeführten Analysen nur zwei recht grobe Ansätze verwendet. Punkt a) wurde hierzu nochmals untergliedert:

- die ermittelte Notation und die korrekte Notation stimmen in der ersten Ziffer überein, oder
- beide Notationen stimmen lediglich in der Hauptklasse überein.

Im ersten Fall wird z.B. die Zuordnung der Notation „WL 100“ zu einem Dokument mit der Notation „WL 150“ als korrekt angesehen. Im zweiten Fall wird darüber hinaus auch die Zuordnung von z.B. der Notation „WN 556“ zu einem Dokument mit der Notation „WN 10“ als korrekte Klassifizierung betrachtet. In beiden Fällen ergeben sich deutliche Ergebnisverbesserungen (siehe Tab. 13): Bei Berücksichtigung der ersten Ziffer werden 67,20 % bzw. 67,40 % der Testdokumente korrekt klassifiziert; bei Berücksichtigung der Hauptklasse beträgt die Übereinstimmung 81,58 % bzw. 81,77 %.

Tab. 13: Anteil korrekt klassifizierter Datensätze in Prozent

	LokNot_ganz	LokNot_kurz
Korrekte Zuordnung	58,07	59,40
… erste Ziffer	67,20	67,40
… nur Hauptklasse	81,58	81,77

9 Interpretation

Die Resultate werden nun im Kontext der in den vorherigen Schritten gewonnenen Erkenntnisse über den Datenbestand interpretiert. Hierdurch lässt sich ihre Aussagekraft einordnen.

9.1 Ergebnisinterpretation

Die einfachste Methode der Klassifizierung neuer Datensätze besteht darin, in der Trainingsphase die größte bekannte Klasse – also die Klasse mit den meisten zugeteilten Dokumenten – zu ermitteln und daraufhin sämtliche zu klassifizierende Dokumente dieser Klasse zuzuordnen. Hierdurch lassen sich bei manchen Datenbeständen schon gute Ergebnisse erzielen. Bei den hier untersuchten Datensätzen ergibt sich bei Anwendung dieser Methode jedoch nur eine Genauigkeit von 2,82 %, da jeder Datensatz der Klasse W 50 zugeordnet wird (siehe Tab. 10 bzw. Tab. 11). In Relation zu diesem sehr geringen Wert sind die hier erreichten

58,07 % bzw. 59,40 % ein deutlich besseres Ergebnis. Im Folgenden wird aufgezeigt, welche Faktoren dieses Ergebnis begünstigt oder ein (noch) besseres Ergebnis verhindert haben.

Abb. 12 verdeutlicht den Zusammenhang zwischen der Klassengröße und einer korrekten Klassifizierung. Eine große Anzahl von Dokumenten in einer Klasse führt im Allgemeinen zu mehr korrekten Klassifizierungen. Nichtsdestotrotz gibt es auch Klassen mit wenigen Dokumenten, die häufig korrekt ermittelt wurden sowie große Klassen, die in den meisten Fällen nicht richtig erkannt wurden. Allein aus der Klassengröße lässt sich also nicht auf die Klassifizierungsqualität schließen, aber sie ermöglicht eine grobe Einordnung.

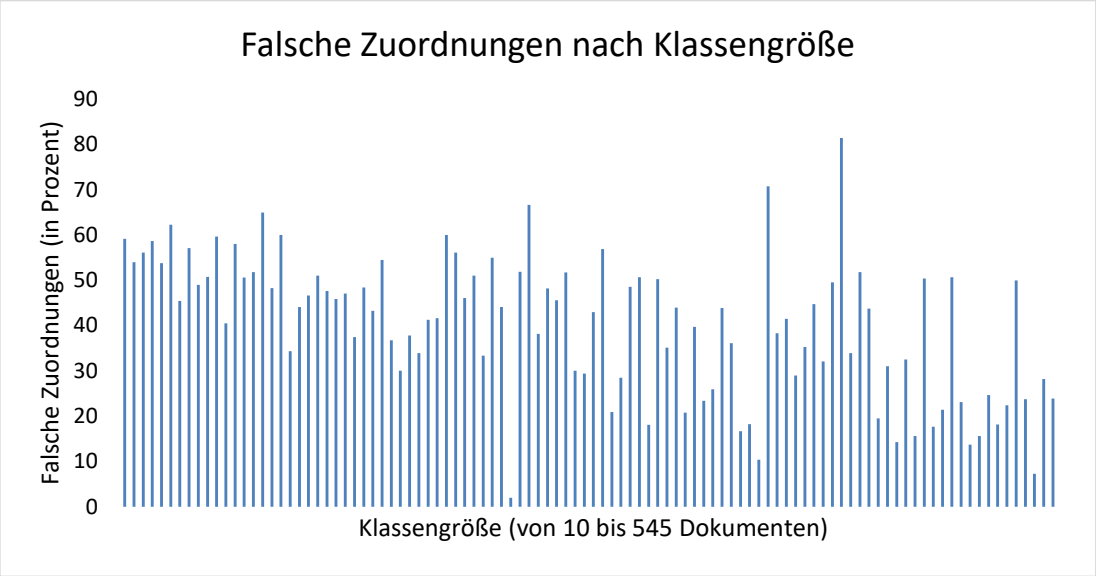

Abb. 12: Falsche Zuordnungen nach Klassengröße

Es besteht ein recht hoher Anteil an Datensätzen, die mit nur einer Klassifikation erschlossen sind (siehe Tab. 12). Gerade bei solchen Dokumenten ist es für ein maschinelles Verfahren schwierig, Unterschiede zwischen einzelnen Notationen zu ermitteln, da nur wenige Informationen über den Datensatz zur Verfügung stehen. Handelt es sich bei der einzigen bekannten Klassifikation zusätzlich um eine eher grobe Klassifikation (wie in diesem Fall die DDC und die BK), dann kann zwischen dieser Klassifikation und der zu lernenden Notation kaum ein Zusammenhang hergestellt werden.

Ein Hauptproblem besteht außerdem darin, dass es vom einzelnen Katalogisierer abhängt, wie ein Datensatz in eine Klassifikation eingeordnet wird. Es gibt in den meisten Fällen kein objektiv beurteilbares „Richtig“ oder „Falsch“, sodass Faktoren wie z.B. Sachkenntnis, Ort, Zeit oder Erfahrung im Katalogisieren eine wichtige Rolle spielen. Ein (semi-)automatisches System basiert auf solchen intellektuell erschlossenen Datensätzen und versucht, Regeln zu definieren, die eine Zuordnung unbekannter Datensätze zu einer Klassifikation ermöglichen.

Von daher können durch solch ein System auch nur diejenigen Informationen gelernt werden, die in den Datensätzen selbst vorhanden sind. Uneinheitliche, fehlerhafte oder sogar widersprüchliche Daten begrenzen daher zwangsläufig das beste zu erreichende Resultat.

9.2 Optimierungsmöglichkeiten

Natürlich sollen die im Rahmen dieser Arbeit durchgeführten Analysen auch kritisch betrachtet werden. Im Folgenden daher eine (sicherlich unvollständige) Aufzählung möglicher Stellen im Data-Mining-Prozess, an denen andere oder ergänzende Entscheidungen getroffen werden können, um möglicherweise ein besseres Ergebnis zu erreichen.

Selektion

Bei der Auswahl der zu betrachtenden Attribute können neben den Klassifikationen weitere, spezifischere Eigenschaften der Datensätze ausgewählt werden. Hierunter fallen z.B. die in anderen Studien oft verwendeten Schlagwörter und Wörter aus dem Titel der in den Katalogdatensätzen repräsentierten Werke.

Datenvorverarbeitung

Bei den Klassifikationen, bei denen in der Datenvorverarbeitung nur die in den Datensätzen erstgenannte Notation verwendet wurde, könnten auch die weiteren Notationen mit einbezogen werden. Hierdurch würde sich ein detaillierteres und differenzierteres Bild der einzelnen Datensätze ergeben, wodurch die Unterscheidung ähnlicher Klassen bzw. nah verwandter Datensätze eventuell einfacher wäre.

Darüber hinaus besteht das Problem der *data sparseness*. Um dieses weiter einzudämmen, könnten bei manchen Klassifikationen durch „Abschneiden“ von spezifischeren Notationsteilen größere – dadurch aber auch thematisch heterogenere – Klassen entstehen. Die Zuordnung zu diesen Klassen ließe sich dann womöglich einfacher gestalten. Allerdings ist hierbei zu berücksichtigen, dass dieser Ansatz von der rein automatischen Sacherschließung wegführt und die Spezifikation der Notationen dann in jedem Fall durch die intellektuelle Arbeit eines Katalogisierers geschehen muss.

Data Mining

Die Verwendung von Abstandsmaßen, um die Ähnlichkeit von zwei Datensätzen zu bestimmen, ist bei nominalen Attributen nur eingeschränkt möglich. Hierbei wird nämlich bei zwei unterschiedlichen Werten der größtmögliche Abstand (also eine vollständige Unähnlichkeit) angenommen, was nicht dem Aufbau hierarchischer Klassifikationen entspricht. Zwei Datensätze, von denen einer z.B. mit „WN 100“ und der andere mit „WN 105“ erschlossen ist,

erhalten bei den üblichen Abstandsmaßen eine Ähnlichkeit von 0. Dabei sind sich diese Datensätze deutlich ähnlicher als z.B. zwei Datensätze mit den Klassen „WN 100“ und „WC 556“. Könnte man diese hierarchischen Beziehungen im Data-Mining-Prozess berücksichtigen, so würde daraus wahrscheinlich eine Ergebnisverbesserung resultieren.

Das verwendete Verfahren geht von einer Gleichwertigkeit der Attribute in den Datensätzen aus. Dies ist genauer betrachtet jedoch nicht der Fall: Das Attribut NLMC ist aussagekräftiger als die anderen Attribute, da es sich hierbei um genau diejenige Klassifikation handelt, die im automatischen Prozess ermittelt werden soll. Durch eine stärkere Gewichtung bekäme dieses Attribut also mehr Einfluss auf das Data-Mining-Verfahren und somit auch auf das Klassifizierungsergebnis.

Weiterhin konnte im Rahmen der vorliegenden Arbeit nur ein Klassifizierungsalgorithmus (KNN) betrachtet werden. Die Verwendung von anderen, möglicherweise besser auf die Ausgangsdaten zugeschnittenen Algorithmen oder von anderen Parametern könnte zu besseren Ergebnissen führen.

10 Fazit

Leitende Fragestellung dieser Arbeit war, inwiefern sich bibliographische Datensätze medizinischer Literatur automatisch klassifizieren lassen, wenn nur die Zuordnung zu anderen Klassifikationen bekannt ist. Hierzu wurde zunächst ein Überblick über bestehende Ansätze und Forschungen in diesem Bereich gegeben. Dieser zeigte bereits eine teilweise unübersichtliche und heterogene Ausgangssituation. Von daher konnte im Rahmen dieser Ausarbeitung nur der Versuch unternommen werden, die Möglichkeiten der automatischen Klassifizierung medizinischer Ressourcen schlaglichtartig zu beleuchten.

Erste Untersuchungen in dieser Richtung weisen darauf hin, dass durchaus zufriedenstellende Ergebnisse zustande gebracht werden können. Natürlich ist dies nur der Anfang: Nähere Analysen der Ausgangsdaten sowie verfeinerte Data-Mining-Methoden werden wahrscheinlich dazu führen, dass die bereits erreichten Resultate noch (deutlich) verbessert werden können. Neben der Berücksichtigung weiterer Attribute (z.B. von MeSH-Termen oder Sachtiteln) stellt insbesondere der Einbezug von weiteren, außerhalb der reinen Katalogdaten liegenden Informations- und Datenquellen (wie z.B. Konkordanzen) hierbei einen vielversprechenden Ansatz dar.

Allerdings haben sich auch Probleme offenbart, die teilweise grundlegender Natur sind und die Möglichkeiten (voll)automatischer Klassifizierung eingrenzen. Hierzu zählt neben der

in Datensätzen produktiver Anwendungen häufig auftretenden *data sparseness* vor allem die Uneinheitlichkeit, mit der Datensätze erschlossen werden:

> „Umsetzbar in eine automatische Lösung sind Aufgaben, die wohldefiniert und in ihren Abläufen und Bestandteilen gut beschreibbar sind, also einer gewissen Regelhaftigkeit folgen. Ob der Vorgang des Indexierens (…) eine solche Aufgabe ist, bleibt wenigstens fraglich. Mater (1990, S. 37) spricht vom Indexieren als einem ‚weitgehend irrationalem Vorgang'. Tatsächlich kann er Untersuchungsergebnisse anführen, die belegen, dass mehrere Indexierer bei der Bearbeitung des selben Textes weniger als 50 Prozent Übereinstimmung (Indexierungskonsistenz) erzielen. Diese Resultate lassen am begriffsorientierten Ansatz des automatischen Indexierens zweifeln, da eine Regelhaftigkeit des Indexierungsvorgangs, die eine notwendige Voraussetzung einer jeden Prozessautomatisierung ist, tatsächlich nicht gegeben bzw. erkennbar zu sein scheint."[119]

Am (auch für die praktische Umsetzung) erfolgversprechendsten scheint der Ansatz der semi-automatischen Klassifizierung zu sein. Wenn die Klassifizierungsaufgabe nicht ausschließlich dem Computer übertragen wird, sondern dieser durch Interaktionen mit dem Katalogisierer während des Klassifizierungsprozesses begleitet wird, werden Genauigkeiten von bis zu 90 % erreicht.[120] Dieser Ansatz konnte im Rahmen dieser Arbeit jedoch nicht weiter ausgeführt werden.

Die hier durchgeführten Untersuchungen sind also nur ein weiterer Baustein im weiten Feld der automatischen Klassifizierung. Dennoch hoffe ich, dass die betrachteten Aspekte bei weiteren Forschungen in dieser Richtung hilfreiche Impulse geben können und vielleicht sogar Grundlagen geschaffen haben, auf denen aufbauend weitere Studien möglich sind. Besonders der nur beiläufig erwähnte Ansatz, Konkordanzen zur Anreicherung der Katalogdaten zu verwenden, scheint mir vielversprechend. So gilt erst einmal weiterhin, was Klaus Lepsky schon 1996 feststellte:

> „Zu denken ist bei einer ‚automatischen' Klassifizierung weniger an eine Klassifizierung im Sinne einer korrekten Notationszuteilung, sondern allenfalls an die Zuordnung des Dokuments zu einer bestimmten fachlichen Thematik. Basis für diese thematische Analyse kann einerseits reichlich vorhandenes Erschließungsvokabular sein, andererseits eine bereits vorhandene klassifikatorische Inhaltserschließung, die über eine Konkordanz genutzt wird."[121]

[119] Nohr (2005), S. 36
[120] Vgl. Wang (2009), S. 2280
[121] Lepsky (1996), S. 71

Literaturverzeichnis

Balakrishnan, Uma (2011): Eine DDC-RVK-Konkordanz. Erste Erkenntnisse aus dem Gebiet „Medizin & Gesundheit“. PowerPoint-Präsentation im Rahmen der 35. Jahrestagung der Gesellschaft für Klassifikation, 1. September 2011, Frankfurt am Main. Online verfügbar unter https://www.gbv.de/Verbundzentrale/Publikationen/2011/pdf/eine-ddc-rvk-konkordanz-erste-erkenntnisse-aus-dem-gebiet-201emedizin-gesundheit, zuletzt geprüft am 23.08.2016.

Balakrishnan, Uma (2013): Das VZG-Projekt „coli-conc“. Brückenbildung zwischen DDC und RVK. PowerPoint-Präsentation im Rahmen des RVK-Projekt Workshop, 20. November 2013, Göttingen. Online verfügbar unter https://www.gbv.de/Verbundzentrale/Publikationen/publikationen-der-vzg-2013/pdf/Balakrishnan_131120_RVK_WS_Konkordanz.pdf, zuletzt geprüft am 23.08.2016.

Balakrishnan, Uma (2015): Cocoda “Colibri Concordance Database” – A mapping tool for library classification schemes. PowerPoint-Präsentation im Rahmen der European Conference on Data Analysis, Gfkl Workshop, 3. September 2015, University of Essex. Online verfügbar unter https://www.gbv.de/Verbundzentrale/Publikationen/publikationen-der-vzg-2015/pdf/Balakrishnan_150903_Cocoda_LIS_Colchester.pdf, zuletzt geprüft am 23.08.2016.

Batley, Sue (2005): Classification in Theory and Practice. Oxford [u.a.]: Chandos Publishing (Chandos Information Professional Series).

Bertram, Jutta (2005): Einführung in die inhaltliche Erschließung. Grundlagen, Methoden, Instrumente. Würzburg: Ergon-Verlag (Content and communication ; Bd. 2).

Borko, Harold; Bernick, Myrna (1963): Automatic Document Classification. In: Journal of the ACM, Jg. 10, H. 2, S. 151-162. Online verfügbar unter http://dx.doi.org/10.1145/321160.321165, zuletzt geprüft am 23.08.2016.

Chakrabarti, Soumen [u.a.] (1998): Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. In: The VLDB Journal, Jg. 7, H. 3, S. 163-178. Online verfügbar unter http://dx.doi.org/10.1007/s007780050061, zuletzt geprüft am 23.08.2016.

Chan, Lois Mai (1999): A guide to the Library of Congress classification. Fifth edition. Englewood, Colorado: Libraries Unlimited (Library and information science text series).

Chan, Lois Mai (2007): Cataloging and Classification. An Introduction. Third edition. Lanham, Maryland [u.a.]: The Scarecrow Press.

Cheng, Patrick T.K.; Wu, Albert K.W. (1995): ACS. An automatic classification system. In: Journal of Information Science, Jg. 21, H. 4, S. 289-299. Online verfügbar unter http://dx.doi.org/10.1177/016555159502100405, zuletzt geprüft am 23.08.2016.

Cleve, Jürgen; Lämmel, Uwe (2016): Data Mining. 2. Aufl. Berlin [u.a.]: De Gruyter (De Gruyter Studium).

Deutsche Nationalbibliothek (o.J.): Dewey-Dezimalklassifikation. Online verfügbar unter http://www.ddc-deutsch.de/Subsites/ddcdeutsch/DE/Home/home_node.html, zuletzt geprüft am 23.08.2016.

Enser, P.G.B. (1985): Automatic classification of book material represented by back-of-the-book index. In: Journal of Documentation, Jg. 41, H. 3, S. 135-155. Online verfügbar unter http://dx.doi.org/10.1108/eb026777, zuletzt geprüft am 23.08.2016.

Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996): From Data Mining to Knowledge Discovery. An Overview. In: Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.): Advances in Knowledge Discovery and Data Mining. Menlo Park [u.a.]: MIT Press, S. 1-34.

Garland, Kathleen (1983): An experiment in automatic hierarchical document classification. In: Information Processing & Management, Jg. 19, H. 3, S. 113-120. Online verfügbar unter http://dx.doi.org/10.1016/0306-4573(83)90064-X, zuletzt geprüft am 23.08.2016.

Gemeinsamer Bibliotheksverbund (2011): Basisklassifikation. 3., erw. Ausg. 2000, letzte Aktualisierung: 4. Ergänzungslieferung Dezember 2011. Online verfügbar unter https://www.gbv.de/vgm/info/mitglieder/02Verbund/01Erschliessung/05Sacherschliessung/05Sacherschliessung_3606.pdf, zuletzt geprüft am 24.08.2016.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data Mining. Concepts and Techniques. Third Edition. Amsterdam [u.a.]: Morgan Kaufmann.

Hermes, Hans-Joachim (1996): Die Konkordanz von Klassifikationen – hat sie eine Chance? In: Hermes, Hans-Joachim; Wätjen, Hans-Joachim (Hrsg.): Erschließen, Suchen, Finden. Vorträge aus den bibliothekarischen Arbeitsgruppen der 19. und 20. Jahrestagungen (Basel 1995/Freiburg 1996) der Gesellschaft für Klassifikation. Oldenburg: Bibliotheks- und Informationssystem der Universität Oldenburg, S. 93-101. Online verfügbar unter http://oops.uni-oldenburg.de/675/73/herwae96.pdf, zuletzt geprüft am 23.08.2016.

Hoyle, W.G. (1973): Automatic indexing and generation of classification systems by algorithm. In: Information Storage and Retrieval, Jg. 9, H. 4, S. 233-242. Online verfügbar unter http://dx.doi.org/10.1016/0020-0271(73)90091-0, zuletzt geprüft am 23.08.2016.

Ishida, Emi (1998): An Experiment of Automatic Classification of Books Using Nippon Decimal Classification. In: Library and Information Science, Jg. 39, S. 31-45. Online verfügbar unter http://lis.mslis.jp/pdf/LIS039031.pdf, zuletzt geprüft am 23.08.2016.

Joorabchi, Arash; Mahdi, Abdulhussain E. (2011): An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. In: Journal of

Information Science, Jg. 37, H. 5, S. 499-514. Online verfügbar unter http://dx.doi.org/10.1177/0165551511417785, zuletzt geprüft am 23.08.2016.

Kar, Gautam; White, Lee J. (1978): A distance measure for automatic document classification by sequential analysis. In: Information Processing & Management, Jg. 14, H. 2, S. 57-69. Online verfügbar unter http://dx.doi.org/10.1016/0306-4573(78)90063-8, zuletzt geprüft am 23.08.2016.

Knorz, Gerhard (1995): Information Retrieval-Anwendungen. In: Zilahi-Szabó, Miklós Géza (Hrsg.): Kleines Lexikon der Informatik und Wirtschaftsinformatik. München: Oldenbourg, S. 244-248.

Larson, Ray R. (1992): Experiments in Automatic Library of Congress Classification. In: Journal of the American Society for Information Science, Jg. 43, H. 2, S. 130-148. Online verfügbar unter http://search.proquest.com/docview/1301247926?accountid=10843, zuletzt geprüft am 23.08.2016.

Lepsky, Klaus (1996): Vom OPAC zum Hyperkatalog: Daten und Indexierung. In: Hermes, Hans-Joachim; Wätjen, Hans-Joachim (Hrsg.): Erschließen, Suchen, Finden. Vorträge aus den bibliothekarischen Arbeitsgruppen der 19. und 20. Jahrestagungen (Basel 1995/Freiburg 1996) der Gesellschaft für Klassifikation. Oldenburg: Bibliotheks- und Informationssystem der Universität Oldenburg, S. 65-74. Online verfügbar unter http://oops.uni-oldenburg.de/675/73/herwae96.pdf, zuletzt geprüft am 23.08.2016.

Library of Congress (o.J.): Library of Congress Classification Outline. Class R – Medicine. Online verfügbar unter https://www.loc.gov/aba/cataloging/classification/lcco/lcco_r.pdf, zuletzt geprüft am 23.08.2016.

Library of Congress (2014): Library of Congress Classification. Online verfügbar unter https://www.loc.gov/catdir/cpso/lcc.html, letzte Änderung am 01.10.2014, zuletzt geprüft am 23.08.2016.

Marcella, Rita; Newton, Robert (1994): A New Manual of Classification. Aldershot [u.a.]: Gower.

Maron, M.E. (1961): Automatic Indexing. An Experimental Inquiry. In: Journal of the ACM, Jg. 8, H. 3, S. 404-417. Online verfügbar unter http://dx.doi.org/10.1145/321075.321084, zuletzt geprüft am 23.08.2016.

Martin, Kristin E.; Mundle, Kavita (2014): Positioning Libraries for a New Bibliographic Universe. A Review of Cataloging and Classification Literature 2011-12. In: Library Resources & Technical Services, Jg. 58, H. 4, S. 233-249. Online verfügbar unter https://journals.ala.org/lrts/article/view/5408/6636, zuletzt geprüft am 23.08.2016.

Mittelbach, Jens; Probst, Michaela (2006): Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung. Strategien für das Bibliothekssystem der Freien Universität Berlin. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; Heft 183). Online verfügbar unter http://www.ib.hu-berlin.de/~kumlau/handreichungen/h183/h183.pdf, zuletzt geprüft am 23.08.2016.

NLM (2016): About the NLM Classification. Online verfügbar unter https://www.nlm.nih.gov/class/nlmclassintro.html, erstellt am 08.10.2002, letzte Änderung am 28.04.2016, zuletzt geprüft am 23.08.2016.

Nöther, Ingo (1994a): Modell einer Konkordanz-Klassifikation für Systematische Kataloge – Teil 1. In: Bibliotheksdienst, Jg. 28, H. 1, S. 15-33. Online verfügbar unter http://dx.doi.org/10.1515/bd.1994.28.1.8, zuletzt geprüft am 23.08.2016.

Nöther, Ingo (1994b): Modell einer Konkordanz-Klassifikation für Systematische Kataloge – Teil 2. In: Bibliotheksdienst, Jg. 28, H. 2, S. 175-187. Online verfügbar unter http://dx.doi.org/10.1515/bd.1994.28.2.155, zuletzt geprüft am 23.08.2016.

Nohr, Holger (2005): Grundlagen der automatischen Indexierung. Ein Lehrbuch. 3., überarb. Aufl. Berlin: Logos-Verl.

Oberhauser, Otto (2005): Automatisches Klassifizieren. Entwicklungsstand, Methodik, Anwendungsbereiche. Frankfurt am Main [u.a.]: Peter Lang (Europäische Hochschulschriften, Reihe XLI Informatik ; Bd. 43).

OCLC (o.J.): DDC 23 Summaries. Online verfügbar unter http://www.oclc.org/content/dam/oclc/dewey/DDC%2023_Summaries.pdf, zuletzt geprüft am 23.08.2016.

Plößnig, Veronika; Steiner, Christoph (2014): Klassifikationen. Konkordanzen, Anreicherungsprojekte und RVK-Datenkorrekturen im Österreichischen Bibliothekenverbund. Ein Update. PowerPoint-Präsentation im Rahmen des RVK-Anwendertreffens 2014. Online verfügbar unter http://rvk.uni-regensburg.de/images/stories/Conf2014/ppt%20plnig_steiner-rvk-bk-12-11-2014.pdf, zuletzt geprüft am 23.08.2016.

Pong, Joanna Yi-Hang [u.a.] (2008): A comparative study of two automatic document classification methods in a library setting. In: Journal of Information Science, Jg. 34, H. 2, S. 213-230. Online verfügbar unter http://dx.doi.org/10.1177/0165551507082592, zuletzt geprüft am 23.08.2016.

Robare, Lori [u.a.] (o.J.): Fundamentals of Library of Congress Classification. Online verfügbar unter https://www.loc.gov/catworkshop/courses/fundamentalslcc/pdf/classify-instr-manual.pdf, zuletzt geprüft am 23.08.2016.

Salton, Gerard; McGill, Michael J. (1987): Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg [u.a.]: McGraw-Hill (McGraw-Hill-Texte).

Schneider, Alexandra (2008): Moderne Retrievalverfahren in klassischen bibliotheksbezogenen Anwendungen. Projekte und Perspektiven. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; Heft 238). Online verfügbar unter http://www.ib.hu-berlin.de/~kumlau/handreichungen/h238/h238.pdf, zuletzt geprüft am 23.08.2016.

Sebastiani, Fabrizio (2002): Machine Learning in Automated Text Categorization. In: ACM Computing Surveys, Jg. 34, H. 1, S. 1-47. Online verfügbar unter http://dx.doi.org/10.1145/505282.505283, zuletzt geprüft am 23.08.2016.

Universitätsbibliothek Regensburg (o.J.): Was ist die RVK? Online verfügbar unter - https://rvk.uni-regensburg.de/2-uncategorised/141-rvk, zuletzt geprüft am 24.08.2016.

Vasuki, Vidya; Cohen, Trevor (2010): Reflective random indexing for semi-automatic indexing of the biomedical literature. In: Journal of Biomedical Informatics, Jg. 43, H. 5, S. 694-700. Online verfügbar unter http://dx.doi.org/10.1016/j.jbi.2010.04.001, zuletzt geprüft am 23.08.2016.

Wang, Jun (2009): An Extensive Study on Automated Dewey Decimal Classification. In: Journal of the American Society for Information Science and Technology, Jg. 60, H. 11, S. 2269-2286. Online verfügbar unter http://dx.doi.org/10.1002/asi.21147, zuletzt geprüft am 23.08.2016.

Witten, Ian H.; Frank, Eibe; Hall, Mark A. (2011): Data Mining. Practical Machine Learning Tools and Techniques. Third edition. Amsterdam [u.a.]: Morgan Kaufmann.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die eingereichte Bachelorarbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Hannover, den 10.12.2016

Andreas Lüschow